Business objective

Your company now sees all the big companies creating original video content and they want to get in on the fun. They have decided to create a new movie studio, but they don't know anything about creating movies. You are charged with exploring what types of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of your company's new movie studio can use to help decide what type of films to create.

Objectives

-
    a. **Identify high-performing movie genres and themes.** Analyze the top 5 years of global box office data to determine which genres and themes generate the highest average revenue and audience ratings.
-
    a. **Analyze audience and market trends.** Track audience preferences and regional performance trends over the past 5–10 years to identify growing market segments.
-
    a. **Evaluate key success factors influencing box office performance.** Quantify how production and marketing factors (budget, runtime, release month, star power, etc.) affect movie profitability.
-
    a. **Provide actionable recommendations for content strategy.** Translate data findings into a business strategy that directs the company's new studio toward commercially viable movie projects.

```
import pandas as pd
import sqlite3

im_conn=sqlite3.connect('./data/im.db')

bom_df = pd.read_csv("./zippedData/bom.movie_gross.csv.gz")
bom_df.head()
```

```
                                          title studio  domestic_gross
\
0                                  Toy Story 3     BV     415000000.0

1                    Alice in Wonderland (2010)     BV     334200000.0

2  Harry Potter and the Deathly Hallows Part 1     WB     296000000.0

3                                     Inception     WB     292600000.0

4                           Shrek Forever After   P/DW     238700000.0
```

```
  foreign_gross  year
0     652000000  2010
1     691300000  2010
2     664300000  2010
3     535700000  2010
4     513900000  2010
```

```
movieinfo_df = pd.read_csv("./zippedData/rt.movie_info.tsv.gz", sep='\
t', compression='gzip')
movieinfo_df.head()
```

```
   id                                               synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                                NaN     NR

                                  genre          director  \
0  Action and Adventure|Classics|Drama  William Friedkin
1    Drama|Science Fiction and Fantasy  David Cronenberg
2    Drama|Musical and Performing Arts    Allison Anders
3          Drama|Mystery and Suspense    Barry Levinson
4                        Drama|Romance    Rodney Bennett

                            writer  theater_date      dvd_date
currency  \
0                   Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
NaN
1     David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
$
2                   Allison Anders  Sep 13, 1996  Apr 18, 2000
NaN
3  Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
NaN
4                     Giles Cooper           NaN           NaN
NaN

  box_office      runtime             studio
0        NaN  104 minutes                NaN
1    600,000  108 minutes  Entertainment One
2        NaN  116 minutes                NaN
3        NaN  128 minutes                NaN
4        NaN  200 minutes                NaN
```

```
reviews_df = pd.read_csv(
    "./zippedData/rt.reviews.tsv.gz",
    sep='\t',
    compression='gzip',
    encoding='latin1'   # or encoding='ISO-8859-1'
```

```
)
reviews_df.head()
```

```
   id                                             review rating
fresh  \
0   3  A distinctly gallows take on contemporary fina...    3/5
fresh
1   3  It's an allegory in search of a meaning that n...    NaN
rotten
2   3  ... life lived in a bubble in financial dealin...    NaN
fresh
3   3  Continuing along a line introduced in last yea...    NaN
fresh
4   3             ... a perverse twist on neorealism...     NaN
fresh

           critic  top_critic         publisher                date
0      PJ Nabarro           0   Patrick Nabarro  November 10, 2018
1  Annalee Newitz           0           io9.com       May 23, 2018
2    Sean Axmaker           0  Stream on Demand    January 4, 2018
3   Daniel Kasman           0              MUBI  November 16, 2017
4             NaN           0      Cinema Scope   October 12, 2017
```

```
tmdb_df = pd.read_csv("./zippedData/tmdb.movies.csv.gz")
tmdb_df.head()
```

```
   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

                                original_title  popularity
release_date  \
0  Harry Potter and the Deathly Hallows: Part 1      33.533   2010-11-
19
1                    How to Train Your Dragon      28.734   2010-03-
26
2                                    Iron Man 2      28.515   2010-05-
07
3                                     Toy Story      28.005   1995-11-
22
4                                     Inception      27.920   2010-07-
16

                                          title  vote_average
vote_count
0  Harry Potter and the Deathly Hallows: Part 1           7.7
10788
```

```
1                        How to Train Your Dragon           7.7
7610
2                                      Iron Man 2           6.8
12368
3                                       Toy Story           7.9
10174
4                                       Inception           8.3
22186
```

```
budgets_df = pd.read_csv("./zippedData/tn.movie_budgets.csv.gz")
budgets_df.head()
```

```
   id  release_date                                          movie  \
0   1  Dec 18, 2009                                         Avatar
1   2  May 20, 2011  Pirates of the Caribbean: On Stranger Tides
2   3   Jun 7, 2019                                   Dark Phoenix
3   4   May 1, 2015                    Avengers: Age of Ultron
4   5  Dec 15, 2017            Star Wars Ep. VIII: The Last Jedi

  production_budget domestic_gross worldwide_gross
0      $425,000,000   $760,507,625  $2,776,345,279
1      $410,600,000   $241,063,875  $1,045,663,875
2      $350,000,000    $42,762,350    $149,762,350
3      $330,600,000   $459,005,868  $1,403,013,963
4      $317,000,000   $620,181,382  $1,316,721,747
```

basic cleaning

```
# check shape and missing values for all datasets

datasets = {
    "BOM": bom_df,
    "Movie Info": movieinfo_df,
    "Reviews": reviews_df,
    "TMDB": tmdb_df,
    "Budgets": budgets_df
}

for name, df in datasets.items():
    print(f"\n{name} dataset: {df.shape[0]} rows, {df.shape[1]}
columns")
    print("Missing values summary:")
    display(df.isnull().sum().sort_values(ascending=True).head(10))


BOM dataset: 3387 rows, 5 columns
Missing values summary:

title                0
year                 0
```

```
studio               5
domestic_gross      28
foreign_gross     1350
dtype: int64


Movie Info dataset: 1560 rows, 12 columns
Missing values summary:

id                  0
rating              3
genre               8
runtime            30
synopsis           62
director          199
theater_date      359
dvd_date          359
writer            449
studio           1066
dtype: int64


Reviews dataset: 54432 rows, 8 columns
Missing values summary:

id                  0
fresh               0
top_critic          0
date                0
publisher         309
critic           2722
review           5563
rating          13517
dtype: int64


TMDB dataset: 26517 rows, 10 columns
Missing values summary:

Unnamed: 0            0
genre_ids             0
id                    0
original_language     0
original_title        0
popularity            0
release_date          0
title                 0
vote_average          0
vote_count            0
dtype: int64
```

```
Budgets dataset: 5782 rows, 6 columns
Missing values summary:

id                   0
release_date         0
movie                0
production_budget    0
domestic_gross       0
worldwide_gross      0
dtype: int64
```

```
#drop duplicates

for name, df in datasets.items():
    df.drop_duplicates(inplace=True)

# this helps you identify if numeric columns like budgets or grosses
are stored as text

for name, df in datasets.items():
    display(df.dtypes)
```

```
title              object
studio             object
domestic_gross    float64
foreign_gross      object
year                int64
dtype: object

id                int64
synopsis         object
rating           object
genre            object
director         object
writer           object
theater_date     object
dvd_date         object
currency         object
box_office       object
runtime          object
studio           object
dtype: object

id              int64
review         object
rating         object
fresh          object
critic         object
top_critic      int64
```

```
publisher     object
date          object
dtype: object
```

```
Unnamed: 0             int64
genre_ids             object
id                     int64
original_language     object
original_title        object
popularity           float64
release_date          object
title                 object
vote_average         float64
vote_count             int64
dtype: object
```

```
id                     int64
release_date          object
movie                 object
production_budget     object
domestic_gross        object
worldwide_gross       object
dtype: object
```

```
# Remove $ and commas, convert to integer

for col in ["production_budget", "domestic_gross", "worldwide_gross"]:
    budgets_df[col] = (budgets_df[col].replace('[\$,]', '',
regex=True).astype(float))

budgets_df[["production_budget", "domestic_gross",
"worldwide_gross"]].head()
```

```
   production_budget  domestic_gross  worldwide_gross
0        425000000.0     760507625.0     2.776345e+09
1        410600000.0     241063875.0     1.045664e+09
2        350000000.0      42762350.0     1.497624e+08
3        330600000.0     459005868.0     1.403014e+09
4        317000000.0     620181382.0     1.316722e+09
```

```
# Make column names consistent across datasets; all lowercase, no
spaces.

bom_df.columns = bom_df.columns.str.lower().str.replace(" ", "_")

movieinfo_df.columns = movieinfo_df.columns.str.lower().str.replace("
", "_")

reviews_df.columns = reviews_df.columns.str.lower().str.replace(" ",
"_")
```

```
tmdb_df.columns = tmdb_df.columns.str.lower().str.replace(" ", "_")

budgets_df.columns = budgets_df.columns.str.lower().str.replace(" ",
"_")


# Create a new database (or connect if exists)
conn = sqlite3.connect("./data/movies_cleaned.db")

# Save each dataframe as a SQL table

bom_df.to_sql("bom_gross", conn, if_exists="replace", index=False)

movieinfo_df.to_sql("rt_movie_info", conn, if_exists="replace",
index=False)

reviews_df.to_sql("rt_reviews", conn, if_exists="replace",
index=False)

tmdb_df.to_sql("tmdb_movies", conn, if_exists="replace", index=False)

budgets_df.to_sql("movie_budgets", conn, if_exists="replace",
index=False)

print("Cleaned datasets loaded into SQLite database successfully!")
```

```
Cleaned datasets loaded into SQLite database successfully!
```

```
# verification of tables in SQLite

pd.read_sql("SELECT name FROM sqlite_master WHERE type='table';",
conn)
```

	name
0	bom_gross
1	rt_movie_info
2	rt_reviews
3	tmdb_movies
4	movie_budgets

```
pd.read_sql("PRAGMA table_info(rt_movie_info);", conn)
```

	cid	name	type	notnull	dflt_value	pk
0	0	id	INTEGER	0	None	0
1	1	synopsis	TEXT	0	None	0
2	2	rating	TEXT	0	None	0
3	3	genre	TEXT	0	None	0
4	4	director	TEXT	0	None	0
5	5	writer	TEXT	0	None	0
6	6	theater_date	TEXT	0	None	0
7	7	dvd_date	TEXT	0	None	0
8	8	currency	TEXT	0	None	0

```
9     9    box_office     TEXT        0       None   0
10   10       runtime     TEXT        0       None   0
11   11        studio     TEXT        0       None   0
```

```
for table in ["bom_gross", "rt_reviews", "tmdb_movies",
"movie_budgets"]:
    print(f"\n{table} columns:")
    display(pd.read_sql(f"PRAGMA table_info({table});", conn))
```

```
bom_gross columns:
```

```
   cid            name     type  notnull dflt_value  pk
0    0           title     TEXT        0       None   0
1    1          studio     TEXT        0       None   0
2    2  domestic_gross     REAL        0       None   0
3    3   foreign_gross     TEXT        0       None   0
4    4            year  INTEGER        0       None   0
```

```
rt_reviews columns:
```

```
   cid        name     type  notnull dflt_value  pk
0    0          id  INTEGER        0       None   0
1    1      review     TEXT        0       None   0
2    2      rating     TEXT        0       None   0
3    3       fresh     TEXT        0       None   0
4    4      critic     TEXT        0       None   0
5    5  top_critic  INTEGER        0       None   0
6    6   publisher     TEXT        0       None   0
7    7        date     TEXT        0       None   0
```

```
tmdb_movies columns:
```

```
   cid               name     type  notnull dflt_value  pk
0    0         unnamed:_0  INTEGER        0       None   0
1    1          genre_ids     TEXT        0       None   0
2    2                 id  INTEGER        0       None   0
3    3  original_language     TEXT        0       None   0
4    4     original_title     TEXT        0       None   0
5    5         popularity     REAL        0       None   0
6    6       release_date     TEXT        0       None   0
7    7              title     TEXT        0       None   0
8    8       vote_average     REAL        0       None   0
9    9         vote_count  INTEGER        0       None   0
```

```
movie_budgets columns:
```

```
   cid               name     type  notnull dflt_value  pk
0    0                 id  INTEGER        0       None   0
```

```
1    1       release_date     TEXT         0       None   0
2    2              movie     TEXT         0       None   0
3    3  production_budget     REAL         0       None   0
4    4     domestic_gross     REAL         0       None   0
5    5    worldwide_gross     REAL         0       None   0
```

OBJECTIVE 1: Identify High-Performing Studios & Genres

We start by finding which studios consistently produce the highest grossing movies.

The `bom_gross` and `movie_budgets` tables are used here.

```
query_studio_performance = """
SELECT
    bg.studio,
    ROUND(AVG(mb.worldwide_gross), 2) AS avg_worldwide_gross,
    COUNT(mb.movie) AS num_movies
FROM movie_budgets mb
JOIN bom_gross bg
    ON mb.movie = bg.title
GROUP BY bg.studio
HAVING num_movies > 3
ORDER BY avg_worldwide_gross DESC
LIMIT 10;
"""

studio_performance_df = pd.read_sql(query_studio_performance, conn)
studio_performance_df
```

```
    studio  avg_worldwide_gross  num_movies
0     P/DW         5.078028e+08          10
1       BV         4.623058e+08          72
2      Fox         2.435983e+08         110
3     Sony         2.378623e+08          74
4     Uni.         2.335837e+08         117
5  WB (NL)         2.308342e+08          37
6       WB         2.175864e+08         102
7     Par.         1.951109e+08          74
8     LG/S         1.230944e+08          31
9     Sum.         1.198865e+08          12
```

OBJECTIVE 2: Analyzing ROI (Profitability)

This calculate each movie's ROI to identify which films and studios achieve the best returns.

```
query_roi = """
SELECT
    mb.movie,
    bg.studio,
    ROUND((mb.worldwide_gross - mb.production_budget) /
```

```
mb.production_budget, 2) AS ROI,
    mb.worldwide_gross,
    mb.production_budget
FROM movie_budgets mb
JOIN bom_gross bg
    ON mb.movie = bg.title
WHERE mb.production_budget > 0
ORDER BY ROI DESC
LIMIT 10;
"""

roi_df = pd.read_sql(query_roi, conn)
roi_df
```

```
                   movie   studio     ROI  worldwide_gross
production_budget
0            The Gallows  WB (NL)  415.56       41656474.0
100000.0
1       The Devil Inside     Par.  100.76      101759490.0
1000000.0
2              Insidious       FD   65.58       99870886.0
1500000.0
3             Unfriended     Uni.   63.36       64364198.0
1000000.0
4  Paranormal Activity 2     Par.   58.17      177512032.0
3000000.0
5                  Split     Uni.   54.79      278964806.0
5000000.0
6                Get Out     Uni.   50.07      255367951.0
5000000.0
7      Chernobyl Diaries       WB   41.41       42411721.0
1000000.0
8  Paranormal Activity 3     Par.   40.41      207039844.0
5000000.0
9              Annabelle  WB (NL)   38.52      256862920.0
6500000.0
```

OBJECTIVE 3: Audience Ratings and Popularity

This shows which genres and types of films receive high audience ratings and votes using TMDb data.

```
query_ratings = """
SELECT
    rmi.genre AS genre,
    ROUND(AVG(tm.vote_average), 2) AS avg_rating,
    COUNT(*) AS num_movies
FROM tmdb_movies tm
JOIN movie_budgets mb
    ON tm.title = mb.movie
```

```
JOIN rt_movie_info rmi
    ON rmi.studio = mb.movie OR rmi.genre IS NOT NULL
GROUP BY rmi.genre
HAVING num_movies > 5
ORDER BY avg_rating DESC
LIMIT 10;
"""

ratings_df = pd.read_sql(query_ratings, conn)
ratings_df
```

```
                                                genre  avg_rating
num_movies
0                                             Western         6.2
11925
1                 Special Interest|Sports and Fitness         6.2
2385
2                                    Special Interest         6.2
2385
3                 Science Fiction and Fantasy|Romance         6.2
2385
4                         Science Fiction and Fantasy         6.2
14310
5  Mystery and Suspense|Science Fiction and Fanta...         6.2
2385
6   Mystery and Suspense|Science Fiction and Fantasy         6.2
7155
7                       Mystery and Suspense|Romance         6.2
2385
8                                Mystery and Suspense         6.2
19080
9       Musical and Performing Arts|Special Interest         6.2
4770
```

OBJECTIVE 4: Movie Performance Trends Over Time

We'll explore whether movie performance has improved or declined over time, based on box office and budgets.

```
query_trends = """
SELECT
    bg.year,
    ROUND(AVG(mb.worldwide_gross), 2) AS avg_gross,
    ROUND(AVG(mb.production_budget), 2) AS avg_budget,
    COUNT(*) AS num_movies
FROM movie_budgets mb
JOIN bom_gross bg
    ON mb.movie = bg.title
GROUP BY bg.year
HAVING num_movies > 5
```

```
ORDER BY bg.year ASC;
"""

trends_df = pd.read_sql(query_trends, conn)
trends_df.head()
```

	year	avg_gross	avg_budget	num_movies
0	2010	1.027785e+08	38876128.53	184
1	2011	1.199072e+08	43302827.38	168
2	2012	1.467869e+08	46617118.06	144
3	2013	1.607725e+08	51617857.14	140
4	2014	1.617406e+08	45311776.35	128

OBJECTIVE 5: Identify Top Performing Movies

This involves finding the most profitable and highest rated movies across all sources.

```
query_best_movies = """
SELECT
    mb.movie,
    bg.studio,
    tm.vote_average AS rating,
    ROUND((mb.worldwide_gross - mb.production_budget) /
mb.production_budget, 2) AS ROI,
    mb.worldwide_gross
FROM movie_budgets mb
JOIN bom_gross bg
    ON mb.movie = bg.title
JOIN tmdb_movies tm
    ON mb.movie = tm.title
WHERE mb.production_budget > 0
ORDER BY ROI DESC, rating DESC
LIMIT 10;
"""

best_movies_df = pd.read_sql(query_best_movies, conn)
best_movies_df
```

	movie	studio	rating	ROI	worldwide_gross
0	The Gallows	WB (NL)	4.8	415.56	41656474.0
1	The Devil Inside	Par.	4.7	100.76	101759490.0
2	Insidious	FD	6.9	65.58	99870886.0
3	Unfriended	Uni.	5.4	63.36	64364198.0
4	Paranormal Activity 2	Par.	5.7	58.17	177512032.0
5	Split	Uni.	7.2	54.79	278964806.0
6	Split	Uni.	5.3	54.79	278964806.0
7	Split	Uni.	5.0	54.79	278964806.0
8	Split	Uni.	4.8	54.79	278964806.0
9	Get Out	Uni.	7.5	50.07	255367951.0

SQL INSIGHTS

Top Studios: The most successful studios generate high worldwide grosses consistently.

ROI Leaders: Low-budget, high-grossing films show strong profit potential.

Ratings: Some genres (from TMDb `genre_ids`) correlate with higher average audience ratings.

Trends: Movie budgets and grosses have shifted over years, showing changing audience interests.

Top Titles: Combining ROI and ratings highlights films that are both profitable and popular.

Data Cleaning

In this section, we do the final cleaning the data from the data sources

1 BOM DF

```
# Load bom_gross
bbom_df = pd.read_sql('''
 SELECT * FROM bom_gross;
''', conn)

bbom_df.head()
```

```
                                          title studio  domestic_gross
\
0                                Toy Story 3     BV     415000000.0

1                  Alice in Wonderland (2010)     BV     334200000.0

2  Harry Potter and the Deathly Hallows Part 1     WB     296000000.0

3                                   Inception     WB     292600000.0

4                           Shrek Forever After   P/DW     238700000.0


  foreign_gross  year
0     652000000  2010
1     691300000  2010
2     664300000  2010
3     535700000  2010
4     513900000  2010
```

```
bbom_df.describe()
```

```
       domestic_gross         year
count    3.359000e+03  3387.000000
mean     2.874585e+07  2013.958075
std      6.698250e+07     2.478141
min      1.000000e+02  2010.000000
```

```
25%      1.200000e+05  2012.000000
50%      1.400000e+06  2014.000000
75%      2.790000e+07  2016.000000
max      9.367000e+08  2018.000000
```

```
bbom_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3387 entries, 0 to 3386
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   title           3387 non-null   object 
 1   studio          3382 non-null   object 
 2   domestic_gross  3359 non-null   float64
 3   foreign_gross   2037 non-null   object 
 4   year            3387 non-null   int64  
dtypes: float64(1), int64(1), object(3)
memory usage: 132.4+ KB
```

```
(2037/3387)*100
```

```
60.14171833480957
```

```
bbom_df["studio"].value_counts()
```

```
IFC          166
Uni.         147
WB           140
Fox          136
Magn.        136
            ...
Swen           1
PalUni         1
IW             1
BSM            1
SMod           1
Name: studio, Length: 257, dtype: int64
```

```
bbom_df[bom_df["studio"].isna()]
```

```
                              title studio  domestic_gross
foreign_gross  year
210   Outside the Law (Hors-la-loi)   None         96900.0
3300000  2010
555         Fireflies in the Garden   None         70600.0
3300000  2011
933           Keith Lemon: The Film   None             NaN
4000000  2012
1862                  Plot for Peace   None          7100.0
None  2014
```

```
2825                Secret Superstar   None              NaN
122000000  2017
```

```
bbom_df[bom_df["studio"]=="NotSpecified"]
```

```
Empty DataFrame
Columns: [title, studio, domestic_gross, foreign_gross, year]
Index: []
```

```
bbom_df["studio"].fillna("NotSpecified", inplace=True)
```

```
bbom_df[bom_df["domestic_gross"].isna() &
bom_df["foreign_gross"].isna()]
```

```
Empty DataFrame
Columns: [title, studio, domestic_gross, foreign_gross, year]
Index: []
```

```
bbom_df[bom_df["domestic_gross"].isna()]
```

```
                                      title        studio
domestic_gross  \
230              It's a Wonderful Afterlife           UTV
NaN
298   Celine: Through the Eyes of the World          Sony
NaN
302                               White Lion         Scre.
NaN
306                         Badmaash Company          Yash
NaN
327                       Aashayein (Wishes)       Relbig.
NaN
537                                    Force          FoxS
NaN
713                         Empire of Silver          NeoC
NaN
871                            Solomon Kane          RTWC
NaN
928                            The Tall Man         Imag.
NaN
933                    Keith Lemon: The Film  NotSpecified
NaN
936                      Lula, Son of Brazil          NYer
NaN
966                           The Cup (2012)          Myr.
NaN
1017                               Dark Tide           WHE
NaN
1079                          The Green Wave            RF
NaN
1268                               22 Bullets         Cdgm.
```

```
NaN
1308              Matru Ki Bijlee Ka Mandola            FIP
NaN
1340                        The Snitch Cartel             PI
NaN
1342             All the Boys Love Mandy Lane           RTWC
NaN
1368                                6 Souls           RTWC
NaN
1659                              Jessabelle            LGF
NaN
1681                              14 Blades           RTWC
NaN
1685          Jack and the Cuckoo-Clock Heart         Shout!
NaN
1739                              Lila Lila          Crnth
NaN
1975           Surprise - Journey To The West             AR
NaN
2392                       Finding Mr. Right 2             CL
NaN
2468                                 Solace            LGP
NaN
2595                                  Viral         W/Dim.
NaN
2825                         Secret Superstar  NotSpecified
NaN

     foreign_gross  year
230        1300000  2010
298         119000  2010
302          99600  2010
306          64400  2010
327           3800  2010
537        4800000  2011
713          19000  2011
871       19600000  2012
928        5200000  2012
933        4000000  2012
936        3800000  2012
966        1800000  2012
1017        432000  2012
1079         70100  2012
1268      21300000  2013
1308       6000000  2013
1340       2100000  2013
1342       1900000  2013
1368        852000  2013
1659       7000000  2014
```

```
1681       3800000  2014
1685       3400000  2014
1739       1100000  2014
1975      49600000  2015
2392     114700000  2016
2468      22400000  2016
2595        552000  2016
2825     122000000  2017
```

```
bbom_df.dropna(subset=["domestic_gross", "foreign_gross"],
inplace=True)
```

We've decided to drop the columns that have missing revenue numbers since we want accurate numbers and filling in with mean or median may inflate or deflate some films hence giving us wrong insights

```
bbom_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2009 entries, 0 to 3353
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           2009 non-null   object
 1   studio          2009 non-null   object
 2   domestic_gross  2009 non-null   float64
 3   foreign_gross   2009 non-null   object
 4   year            2009 non-null   int64
dtypes: float64(1), int64(1), object(3)
memory usage: 94.2+ KB
```

```
#Connecting to cleaned movies
cleaned_conn = sqlite3.connect("./cleaned_data/movies.db")

# Saving to cleaned database
bbom_df.to_sql("bom_gross", cleaned_conn, if_exists="replace",
index=False)
```

2. RT Movie Info

```
# Load RT Movie info
bmovieinfo_df = pd.read_sql('''
 SELECT * FROM rt_movie_info
''', conn)

bmovieinfo_df.head()
```

```
   id                                           synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
```

```
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                                None     NR

                                  genre          director  \
0  Action and Adventure|Classics|Drama  William Friedkin
1    Drama|Science Fiction and Fantasy  David Cronenberg
2    Drama|Musical and Performing Arts    Allison Anders
3           Drama|Mystery and Suspense    Barry Levinson
4                       Drama|Romance    Rodney Bennett

                            writer  theater_date      dvd_date
currency  \
0                   Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
None
1     David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
$
2                   Allison Anders  Sep 13, 1996  Apr 18, 2000
None
3  Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
None
4                     Giles Cooper          None          None
None

  box_office      runtime             studio
0       None  104 minutes               None
1    600,000  108 minutes  Entertainment One
2       None  116 minutes               None
3       None  128 minutes               None
4       None  200 minutes               None

#Describe
bmovieinfo_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1560 entries, 0 to 1559
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   id            1560 non-null   int64
 1   synopsis      1498 non-null   object
 2   rating        1557 non-null   object
 3   genre         1552 non-null   object
 4   director      1361 non-null   object
 5   writer        1111 non-null   object
 6   theater_date  1201 non-null   object
 7   dvd_date      1201 non-null   object
 8   currency      340 non-null    object
 9   box_office    340 non-null    object
 10  runtime       1530 non-null   object
```

```
 11  studio        494 non-null    object
dtypes: int64(1), object(11)
memory usage: 146.4+ KB
```

```
# Checking for all the currencies in the table
bmovieinfo_df["currency"].value_counts()
```

```
$    340
Name: currency, dtype: int64
```

We are dropping currency columns since they are all in dollars hence it is a redudant column

```
bmovieinfo_df.drop(columns="currency", inplace=True)
```

```
bmovieinfo_df
```

```
        id                                               synopsis
rating  \
0        1  This gritty, fast-paced, and innovative police...      R

1        3  New York City, not-too-distant-future: Eric Pa...      R

2        5  Illeana Douglas delivers a superb performance ...      R

3        6  Michael Douglas runs afoul of a treacherous su...      R

4        7                                               None     NR

...    ...                                                ...    ...

1555  1996  Forget terrorists or hijackers -- there's a ha...      R

1556  1997  The popular Saturday Night Live sketch was exp...     PG

1557  1998  Based on a novel by Richard Powell, when the l...      G

1558  1999  The Sandlot is a coming-of-age story about a g...     PG

1559  2000  Suspended from the force, Paris cop Hubert is ...      R


                                                  genre
director  \
0                  Action and Adventure|Classics|Drama    William
Friedkin
1                    Drama|Science Fiction and Fantasy    David
Cronenberg
2                    Drama|Musical and Performing Arts      Allison
Anders
3                           Drama|Mystery and Suspense      Barry
Levinson
```

```
4                                                 Drama|Romance      Rodney
Bennett
...                                                         ...
...
1555   Action and Adventure|Horror|Mystery and Suspense
None
1556                  Comedy|Science Fiction and Fantasy        Steve
Barron
1557  Classics|Comedy|Drama|Musical and Performing Arts      Gordon
Douglas
1558    Comedy|Drama|Kids and Family|Sports and Fitness  David Mickey
Evans
1559  Action and Adventure|Art House and Internation...
None

                                                    writer
theater_date  \
0                                          Ernest Tidyman   Oct 9, 1971

1                          David Cronenberg|Don DeLillo  Aug 17, 2012

2                                          Allison Anders  Sep 13, 1996

3                        Paul Attanasio|Michael Crichton   Dec 9, 1994

4                                            Giles Cooper          None

...                                                   ...           ...

1555                                                None  Aug 18, 2006

1556  Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner  Jul 23, 1993

1557                                                None   Jan 1, 1962

1558                    David Mickey Evans|Robert Gunter   Apr 1, 1993

1559                                          Luc Besson  Sep 27, 2001


          dvd_date  box_office      runtime             studio
0     Sep 25, 2001        None  104 minutes               None
1      Jan 1, 2013     600,000  108 minutes  Entertainment One
2     Apr 18, 2000        None  116 minutes               None
3     Aug 27, 1997        None  128 minutes               None
4             None        None  200 minutes               None
...            ...         ...         ...                ...
1555   Jan 2, 2007  33,886,034  106 minutes    New Line Cinema
1556  Apr 17, 2001        None   88 minutes  Paramount Vantage
1557  May 11, 2004        None  111 minutes               None
```

```
1558  Jan 29, 2002        None  101 minutes               None
1559  Feb 11, 2003        None   94 minutes  Columbia Pictures

[1560 rows x 11 columns]
```

```
bmovieinfo_df[bmovieinfo_df["box_office"].isna() == False]
["box_office"]
```

```
1           600,000
6        41,032,915
7           224,114
8           134,904
15        1,039,869
            ...
1541     25,335,935
1542      1,416,189
1545         59,371
1546        794,306
1555     33,886,034
Name: box_office, Length: 340, dtype: object
```

```
# Marking missing box_office_missing since it's an important column
even though most values are missing
# This makes it easier to filter out filled in values
bmovieinfo_df['box_office_missing'] =
bmovieinfo_df['box_office'].isna()

#bmovieinfo_df.drop(columns="box_office_missing", inplace=True)

bmovieinfo_df.head(10)
```

```
   id                                              synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                               None     NR
5   8  The year is 1942. As the Allies unite overseas...     PG
6  10  Some cast and crew from NBC's highly acclaimed...  PG-13
7  13  Stewart Kane, an Irishman living in the Austra...      R
8  14  "Love Ranch" is a bittersweet love story that ...      R
9  15  When a diamond expedition in the Congo is lost...  PG-13

                                            genre          director
\
0                Action and Adventure|Classics|Drama  William Friedkin

1                  Drama|Science Fiction and Fantasy  David Cronenberg

2                  Drama|Musical and Performing Arts    Allison Anders
```

```
3                          Drama|Mystery and Suspense    Barry Levinson
4                                      Drama|Romance    Rodney Bennett
5                                Drama|Kids and Family       Jay Russell
6                                              Comedy       Jake Kasdan
7                                               Drama      Ray Lawrence
8                                               Drama   Taylor Hackford
9  Action and Adventure|Mystery and Suspense|Scie...    Frank Marshall

                                writer  theater_date      dvd_date
box_office  \
0                     Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
None
1      David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
600,000
2                     Allison Anders  Sep 13, 1996  Apr 18, 2000
None
3   Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
None
4                       Giles Cooper          None          None
None
5                    Gail Gilchriest   Mar 3, 2000  Jul 11, 2000
None
6                         Mike White  Jan 11, 2002  Jun 18, 2002
41,032,915
7  Raymond Carver|Beatrix Christian  Apr 27, 2006   Oct 2, 2007
224,114
8                      Mark Jacobson  Jun 30, 2010   Nov 9, 2010
134,904
9               John Patrick Shanley   Jun 9, 1995  Jul 27, 1999
None

       runtime                   studio  box_office_missing
0  104 minutes                     None               True
1  108 minutes       Entertainment One              False
2  116 minutes                     None               True
3  128 minutes                     None               True
4  200 minutes                     None               True
5   95 minutes   Warner Bros. Pictures               True
6   82 minutes      Paramount Pictures              False
7  123 minutes  Sony Pictures Classics              False
8  117 minutes                     None              False
9  108 minutes                     None               True
```

```
# Remove commas
bmovieinfo_df['box_office'] = (
    bmovieinfo_df['box_office']
    .replace('None', pd.NA)
    .str.replace(',', '', regex=True)  # remove commas
)

# Convert box_office to numeric data type
bmovieinfo_df["box_office"] =
pd.to_numeric(bmovieinfo_df["box_office"], errors="coerce")

bmovieinfo_df.head(10)
```

```
   id                                           synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                               None     NR
5   8  The year is 1942. As the Allies unite overseas...     PG
6  10  Some cast and crew from NBC's highly acclaimed...  PG-13
7  13  Stewart Kane, an Irishman living in the Austra...      R
8  14  "Love Ranch" is a bittersweet love story that ...      R
9  15  When a diamond expedition in the Congo is lost...  PG-13

                                          genre          director
\
0               Action and Adventure|Classics|Drama  William Friedkin

1                 Drama|Science Fiction and Fantasy  David Cronenberg

2                 Drama|Musical and Performing Arts    Allison Anders

3                       Drama|Mystery and Suspense    Barry Levinson

4                                   Drama|Romance    Rodney Bennett

5                            Drama|Kids and Family       Jay Russell

6                                          Comedy       Jake Kasdan

7                                           Drama      Ray Lawrence

8                                           Drama   Taylor Hackford

9  Action and Adventure|Mystery and Suspense|Scie...    Frank Marshall


                           writer  theater_date      dvd_date
box_office  \
0                   Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
```

```
NaN
1      David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
600000.0
2                     Allison Anders  Sep 13, 1996  Apr 18, 2000
NaN
3   Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
NaN
4                       Giles Cooper          None          None
NaN
5                    Gail Gilchriest   Mar 3, 2000  Jul 11, 2000
NaN
6                         Mike White  Jan 11, 2002  Jun 18, 2002
41032915.0
7  Raymond Carver|Beatrix Christian  Apr 27, 2006   Oct 2, 2007
224114.0
8                      Mark Jacobson  Jun 30, 2010   Nov 9, 2010
134904.0
9               John Patrick Shanley   Jun 9, 1995  Jul 27, 1999
NaN
```

	runtime	studio	box_office_missing
0	104 minutes	None	True
1	108 minutes	Entertainment One	False
2	116 minutes	None	True
3	128 minutes	None	True
4	200 minutes	None	True
5	95 minutes	Warner Bros. Pictures	True
6	82 minutes	Paramount Pictures	False
7	123 minutes	Sony Pictures Classics	False
8	117 minutes	None	False
9	108 minutes	None	True

```
bmovieinfo_df.describe()
```

	id	box_office
count	1560.000000	3.400000e+02
mean	1007.303846	3.790601e+07
std	579.164527	5.749159e+07
min	1.000000	3.630000e+02
25%	504.750000	1.905152e+06
50%	1007.500000	1.414105e+07
75%	1503.250000	4.482524e+07
max	2000.000000	3.680000e+08

```
rt_info_bo_median = bmovieinfo_df["box_office"].median()
rt_info_bo_median
```

```
14141054.5
```

```
# Filling missing box office values with the box_office column median
bmovieinfo_df["box_office"].fillna(rt_info_bo_median,inplace=True)

bmovieinfo_df.head(10)
```

```
   id                                             synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                               None     NR
5   8  The year is 1942. As the Allies unite overseas...     PG
6  10  Some cast and crew from NBC's highly acclaimed...  PG-13
7  13  Stewart Kane, an Irishman living in the Austra...      R
8  14  "Love Ranch" is a bittersweet love story that ...      R
9  15  When a diamond expedition in the Congo is lost...  PG-13

                                                genre          director
\
0                Action and Adventure|Classics|Drama  William Friedkin

1                  Drama|Science Fiction and Fantasy  David Cronenberg

2                  Drama|Musical and Performing Arts    Allison Anders

3                          Drama|Mystery and Suspense    Barry Levinson

4                                       Drama|Romance    Rodney Bennett

5                               Drama|Kids and Family       Jay Russell

6                                             Comedy       Jake Kasdan

7                                              Drama      Ray Lawrence

8                                              Drama   Taylor Hackford

9  Action and Adventure|Mystery and Suspense|Scie...    Frank Marshall


                               writer  theater_date      dvd_date
box_office  \
0                     Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
14141054.5
1      David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
600000.0
2                     Allison Anders  Sep 13, 1996  Apr 18, 2000
14141054.5
3   Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
14141054.5
4                       Giles Cooper          None          None
```

```
14141054.5
5                  Gail Gilchriest   Mar 3, 2000  Jul 11, 2000
14141054.5
6                        Mike White  Jan 11, 2002  Jun 18, 2002
41032915.0
7  Raymond Carver|Beatrix Christian  Apr 27, 2006   Oct 2, 2007
224114.0
8                     Mark Jacobson  Jun 30, 2010   Nov 9, 2010
134904.0
9              John Patrick Shanley   Jun 9, 1995  Jul 27, 1999
14141054.5

       runtime                   studio  box_office_missing
0  104 minutes                     None               True
1  108 minutes       Entertainment One               False
2  116 minutes                     None               True
3  128 minutes                     None               True
4  200 minutes                     None               True
5   95 minutes   Warner Bros. Pictures               True
6   82 minutes      Paramount Pictures               False
7  123 minutes  Sony Pictures Classics               False
8  117 minutes                     None               False
9  108 minutes                     None               True
```

We have handled the box office column by marking missing values as missing and then filling them with median

```
bmovieinfo_df[bmovieinfo_df["director"].isna()]

        id                                              synopsis
rating  \
10      17                                                  None   None

11      18  In 1979, Bill Viola and Frank Caliguri dreamed...     NR

12      19  While Microsoft may be the biggest software co...     NR

16      23  A fictional film set in the alluring world of ...      R

20      27                                                  None     NR

...    ...                                                   ...    ...

1543  1982                                                  None   None

1546  1986  Aki Kaurismaki's The Man Without a Past opens ...     PG

1549  1989  Hungarian Rhapsody (Magyar Rapszodia) is the f...     NR

1555  1996  Forget terrorists or hijackers -- there's a ha...      R
```

```
1559  2000  Suspended from the force, Paris cop Hubert is ...      R


                                                      genre director  \
10                                                     None     None
11                                                Documentary     None
12                          Documentary|Special Interest     None
16                                                 Drama     None
20                           Musical and Performing Arts     None
...                                                   ...      ...
1543                                                  None     None
1546           Art House and International|Comedy|Drama     None
1549                  Art House and International|Drama     None
1555   Action and Adventure|Horror|Mystery and Suspense     None
1559  Action and Adventure|Art House and Internation...     None

            writer  theater_date      dvd_date  box_office
runtime  \
10            None          None          None  14141054.5
None
11    Robert Zullo          None          None  14141054.5
None
12            None  Aug 23, 2002  Sep 30, 2003  14141054.5   90
minutes
16            None  Dec 20, 2013  Mar 18, 2014  99165609.0  129
minutes
20            None          None          None  14141054.5
None
...            ...           ...           ...         ...          ..
.
1543          None          None          None  14141054.5
None
1546          None  Aug 30, 2002   Oct 7, 2003    794306.0   97
minutes
1549          None          None          None  14141054.5  101
minutes
1555          None  Aug 18, 2006   Jan 2, 2007  33886034.0  106
minutes
1559   Luc Besson  Sep 27, 2001  Feb 11, 2003  14141054.5   94
minutes

                         studio  box_office_missing
10                         None               True
11    Showtime Documentary Films               True
12         Seventh Art Releasing               True
16                 Sony Pictures              False
20                         None               True
...                         ...                ...
1543                       None               True
```

```
1546                          None                False
1549                          None                 True
1555              New Line Cinema                False
1559            Columbia Pictures                 True

[199 rows x 12 columns]
```

```
bmovieinfo_df["runtime"].head()
```

```
0    104 minutes
1    108 minutes
2    116 minutes
3    128 minutes
4    200 minutes
Name: runtime, dtype: object
```

```
bmovieinfo_df['runtime'].str.replace('minutes', '', regex=True)  # remove commas
```

```
0       104
1       108
2       116
3       128
4       200
       ...
1555    106
1556     88
1557    111
1558    101
1559     94
Name: runtime, Length: 1560, dtype: object
```

```
bmovieinfo_df['runtime'] = (
    bmovieinfo_df['runtime']
    .str.replace('minutes', '', regex=True)  # remove commas
)
```

bmovieinfo_df["box_office"] = pd.to_numeric(bmovieinfo_df["box_office"], errors="coerce")

```
bmovieinfo_df["runtime"] = pd.to_numeric(bmovieinfo_df["runtime"], errors="coerce")

bmovieinfo_df[bmovieinfo_df['runtime'].isna()]
```

```
        id                                                synopsis
rating  \
10      17                                                    None   None

11      18  In 1979, Bill Viola and Frank Caliguri dreamed...     NR

20      27                                                    None     NR
```

102	131	No Sesame. All Street. THE HAPPYTIME MURDERS i...	R
131	167	None	None
195	258	None	NR
200	265	Wakeboarding is a sport of ever-increasing pop...	NR
434	567	Now graduated from college and out in the real...	PG-13
486	636	None	NR
516	676	The Hill would have made a terrific Samuel Ful...	NR
536	699	None	NR
555	724	None	NR
573	743	None	NR
579	749	Jalaibee is a Tale of two friends Billu & ...	NR
750	968	Arnab a typical Bengali man is making his way ...	NR
829	1074	None	NR
921	1192	None	NR
923	1195	None	NR
976	1267	None	NR
1023	1325	From the outer reaches of space to the small-t...	R
1078	1389	From Ron Shelton, writer/director of Tin Cup a...	PG-13
1126	1451	A group of scientific researchers on a space s...	NR
1143	1473	In her return we find Red Sonja, a young girl ...	NR
1201	1541	In the heatwarming live action adventure "Disn...	PG
1342	1736	None	NR
1369	1768	None	NR
1412	1821	None	NR
1487	1913	None	NR
1499	1931	Mark Felt - The Man Who Brought Down the White...	PG-13

```
1543  1982                                                  None   None


                                                    genre
director  \
10                                                   None
None
11                                             Documentary
None
20                              Musical and Performing Arts
None
102                             Action and Adventure|Comedy      Brian
Henson
131                                                  None
None
195                       Art House and International|Drama
None
200                     Special Interest|Sports and Fitness
None
434                                                Comedy         Trish
Sie
486                                     Special Interest    Andreas
Morell
516                              Action and Adventure|Drama     Robert
Iscove
536                                                Drama      Chris
Menges
555   Art House and International|Drama|Sports and F...
None
573                                                Drama
None
579   Action and Adventure|Art House and Internation...      Yasir
Jaswal
750   Art House and International|Horror|Mystery and...        Prosit
Roy
829                                                Drama      Craig
Brewer
921                                               Horror
None
923                                                Drama
None
976   Action and Adventure|Kids and Family|Science F...       Guy
Ritchie
1023  Action and Adventure|Horror|Science Fiction an...       Shane
Black
1078                            Action and Adventure|Comedy       Ron
Shelton
1126   Mystery and Suspense|Science Fiction and Fantasy       Julius
Onah
```

```
1143                          Action and Adventure|Horror
None
1201        Action and Adventure|Comedy|Kids and Family      Marc
Forster
1342                                                 None
None
1369                                          Documentary   Lina
Mannheimer
1412           Art House and International|Comedy|Drama  Roman
Bondarchuk
1487                                Action and Adventure
None
1499                                                Drama   Peter
Landesman
1543                                                 None
None

                                              writer  theater_date
dvd_date  \
10                                              None          None
None
11                                    Robert Zullo          None
None
20                                              None          None
None
102             Todd Berger|Dee Austin Robertson  Aug 24, 2018   Dec
4, 2018
131                                             None          None
None
195                                             None          None
None
200                                             None          None
None
434                          Mike White|Kay Cannon  Dec 22, 2017  Mar
20, 2018
486                                             None          None
None
516                                             None          None
None
536                                     Shawn Slovo          None
None
555                                             None          None
None
573                                             None          None
None
579                                             None          None
None
750   Abhishek Bannerjee|Prosit Roy|Rajat Kapoor          None
None
```

```
829                                              None          None
None
921                                              None          None
None
923                                              None          None
None
976                          John August|Guy Ritchie          None
None
1023                         Shane Black|Fred Dekker  Sep 14, 2018  Nov
27, 2018
1078                                    Ron Shelton   Dec 8, 2017  Feb
27, 2018
1126                                     Oren Uziel   Feb 4, 2018   Feb
5, 2019
1143                                             None          None
None
1201                                Alex Ross Perry   Aug 3, 2018   Nov
6, 2018
1342                                             None          None
None
1369                                             None          None
None
1412                                             None          None
None
1487                                             None          None
None
1499                                Peter Landesman  Sep 29, 2017   Jan
9, 2018
1543                                             None          None
None

       box_office  runtime                        studio
box_office_missing
10     14141054.5      NaN                          None
True
11     14141054.5      NaN  Showtime Documentary Films
True
20     14141054.5      NaN                          None
True
102    14141054.5      NaN                       STXfilms
True
131    14141054.5      NaN                          None
True
195    14141054.5      NaN                          None
True
200    14141054.5      NaN                          None
True
434   104880310.0      NaN                          None
False
```

```
486    14141054.5      NaN                          None
True
516    14141054.5      NaN                          None
True
536    14141054.5      NaN                          None
True
555    14141054.5      NaN                          None
True
573    14141054.5      NaN                          None
True
579    14141054.5      NaN                          None
True
750    14141054.5      NaN                          None
True
829    14141054.5      NaN                          None
True
921    14141054.5      NaN                          None
True
923    14141054.5      NaN                          None
True
976    14141054.5      NaN                          None
True
1023   14141054.5      NaN                          None
True
1078   14141054.5      NaN          Broad Green Pictures
True
1126   14141054.5      NaN                          None
True
1143   14141054.5      NaN                          None
True
1201   14141054.5      NaN          Walt Disney Pictures
True
1342   14141054.5      NaN                          None
True
1369   14141054.5      NaN                          None
True
1412   14141054.5      NaN                          None
True
1487   14141054.5      NaN                          None
True
1499     766428.0      NaN       Sony Pictures Classics
False
1543   14141054.5      NaN                          None
True

bmovieinfo_df["runtime"].describe()

count    1530.000000
mean      103.967974
std        24.642392
```

```
min          5.000000
25%         91.000000
50%        100.000000
75%        114.000000
max        358.000000
Name: runtime, dtype: float64
```

```
bmvinf_run_mean = bmovieinfo_df['runtime'].mean()
bmvinf_run_mean
```

```
103.96797385620916
```

```
bmovieinfo_df['runtime'].fillna(bmvinf_run_mean, inplace = True)
```

```
bmovieinfo_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1560 entries, 0 to 1559
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   id                  1560 non-null   int64  
 1   synopsis            1498 non-null   object 
 2   rating              1557 non-null   object 
 3   genre               1552 non-null   object 
 4   director            1361 non-null   object 
 5   writer              1111 non-null   object 
 6   theater_date        1201 non-null   object 
 7   dvd_date            1201 non-null   object 
 8   box_office          1560 non-null   float64
 9   runtime             1560 non-null   float64
 10  studio              494 non-null    object 
 11  box_office_missing  1560 non-null   bool   
dtypes: bool(1), float64(2), int64(1), object(8)
memory usage: 135.7+ KB
```

```
# Saving to cleaned database
bmovieinfo_df.to_sql("rt_movie_info", cleaned_conn,
if_exists="replace", index=False)
```

3. RT Reviews

```
# Loading RT Reviews
breview_df = pd.read_sql('''
    SELECT * FROM rt_reviews;
''', conn)

breview_df.head(10)
```

```
   id                                             review rating
fresh  \
0   3  A distinctly gallows take on contemporary fina...    3/5
fresh
1   3  It's an allegory in search of a meaning that n...   None
rotten
2   3  ... life lived in a bubble in financial dealin...   None
fresh
3   3  Continuing along a line introduced in last yea...   None
fresh
4   3              ... a perverse twist on neorealism...    None
fresh
5   3  ... Cronenberg's Cosmopolis expresses somethin...   None
fresh
6   3  Quickly grows repetitive and tiresome, meander...      C
rotten
7   3  Cronenberg is not a director to be daunted by ...    2/5
rotten
8   3  Cronenberg's cold, exacting precision and emot...   None
fresh
9   3  Over and above its topical urgency or the bit ...   None
fresh

             critic  top_critic           publisher                  date

0       PJ Nabarro           0     Patrick Nabarro   November 10, 2018

1   Annalee Newitz           0             io9.com        May 23, 2018

2     Sean Axmaker           0    Stream on Demand     January 4, 2018

3    Daniel Kasman           0                MUBI   November 16, 2017

4             None           0        Cinema Scope    October 12, 2017

5  Michelle Orange           0    Capital New York  September 11, 2017

6   Eric D. Snider           0     EricDSnider.com       July 17, 2013

7     Matt Kelemen           0  Las Vegas CityLife      April 21, 2013

8     Sean Axmaker           0       Parallax View      March 24, 2013

9     Kong Rithdee           0        Bangkok Post       March 4, 2013


breview_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54423 entries, 0 to 54422
Data columns (total 8 columns):
```

```
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   id          54423 non-null  int64
 1   review      48867 non-null  object
 2   rating      40907 non-null  object
 3   fresh       54423 non-null  object
 4   critic      51710 non-null  object
 5   top_critic  54423 non-null  int64
 6   publisher   54114 non-null  object
 7   date        54423 non-null  object
dtypes: int64(2), object(6)
memory usage: 3.3+ MB
```

I'm dropping rows with missing ratings since they are of no use to as

```
breview_df.dropna(subset=["rating"], inplace=True)

breview_df
```

```
         id                                             review rating
fresh  \
0         3  A distinctly gallows take on contemporary fina...    3/5
fresh
6         3  Quickly grows repetitive and tiresome, meander...      C
rotten
7         3  Cronenberg is not a director to be daunted by ...    2/5
rotten
11        3  While not one of Cronenberg's stronger films, ...     B-
fresh
12        3  Robert Pattinson works mighty hard to make Cos...    2/4
rotten
...     ...                                                ...    ...
...
54415  2000  Dawdles and drags when it should pop; it doesn...  1.5/5
rotten
54419  2000                                               None    1/5
rotten
54420  2000                                               None    2/5
rotten
54421  2000                                               None  2.5/5
rotten
54422  2000                                               None    3/5
fresh

                   critic  top_critic           publisher
date
0             PJ Nabarro           0     Patrick Nabarro   November
10, 2018
6         Eric D. Snider           0     EricDSnider.com       July
```

```
17, 2013
7          Matt Kelemen           0  Las Vegas CityLife      April
21, 2013
11         Emanuel Levy           0     EmanuelLevy.Com    February
3, 2013
12       Christian Toto           0       Big Hollywood    January
15, 2013
...                  ...         ...                 ...
...
54415     Manohla Dargis           1   Los Angeles Times  September
26, 2002
54419  Michael Szymanski           0          Zap2it.com  September
21, 2005
54420       Emanuel Levy           0     EmanuelLevy.Com       July
17, 2005
54421   Christopher Null           0      Filmcritic.com   September
7, 2003
54422    Nicolas Lacroix           0        Showbizz.net   November
12, 2002

[40907 rows x 8 columns]
```

```
breview_df['rating'].unique()
```

```
array(['3/5', 'C', '2/5', 'B-', '2/4', 'B', '3/4', '4/5', '4/4',
'6/10',
       '1/4', '8', '2.5/4', '4/10', '2.0/5', '3/10', '7/10', 'A-',
'5/5',
       'F', '3.5/4', 'D+', '1.5/4', '3.5/5', '8/10', 'B+', '9/10',
       '2.5/5', '7.5/10', '5.5/10', 'C-', '1.5/5', '1/5', '5/10',
'C+',
       '0/5', '6', '0.5/4', 'D', '3.1/5', '3/6', '4.5/5', '0/4',
'2/10',
       'D-', '7', '1/10', '3', 'A+', 'A', '4.0/4', '9.5/10', '2.5',
       '2.1/2', '6.5/10', '3.7/5', '8.4/10', '9', '1', '7.2/10',
'2.2/5',
       '0.5/10', '5', '0', '2', '4.5', '7.7', '5.0/5', '8.5/10',
'3.0/5',
       '0.5/5', '1.5/10', '3.0/4', '2.3/10', '4.5/10', '4/6', '3.5',
       '8.6/10', '6/8', '2.0/4', '2.7', '4.2/10', '5.8', '4',
'7.1/10',
       '5/4', 'N', '3.5/10', '5.8/10', 'R', '4.0/5', '0/10', '5.0/10',
       '5.9/10', '2.4/5', '1.9/5', '4.9', '7.4/10', '1.5', '2.3/4',
       '8.8/10', '4.0/10', '2.2', '3.8/10', '6.8/10', '7.3', '7.0/10',
       '3.2', '4.2', '8.4', '5.5/5', '6.3/10', '7.6/10', '8.1/10',
       '3.6/5', '2/6', '7.7/10', '1.8', '8.9/10', '8.9', '8.2/10',
       '8.3/10', '2.6/6', '4.1/10', '2.5/10', 'F+', '6.0/10', '1.0/4',
       '7.9/10', '8.7/10', '4.3/10', '9.6/10', '9.0/10', '4.0', '1.7',
       '7.9', '6.7', '8.0/10', '9.2/10', '5.2', '5.9', '3.7', '4.7',
       '6.2/10', '1/6', '8.2', '2.6/5', '3.4', '9.7', '3.3/5',
```

```
'3.8/5',
       '1/2', '7.4', '4.8', '1.6/5', '2/2', '1-5', '1.0', '4.3/5',
'5/6',
       '9.2', '2.7/5', '4.9/10', '3.0', '3.1', '7.8/10', 'F-',
'2.3/5',
       '3.0/10', '3/2', '7.8', '4.2/5', '9.0', '7.3/10', '4.4/5',
       '6.9/10', '0/6', 'T', '6.2', '3.3', '9.8', '8.5', '1.0/5',
'4.1',
       '7.1', '3 1/2'], dtype=object)
```

```
# Saving to cleaned database
breview_df.to_sql("rt_reviews", cleaned_conn, if_exists="replace",
index=False)
```

5. TMDB

```
    # Loading the tmdb database
btmdb_df = pd.read_sql('''
    SELECT * FROM tmdb_movies;
''',conn)

btmdb_df
```

```
       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

                                     original_title  popularity
release_date  \
0      Harry Potter and the Deathly Hallows: Part 1      33.533
2010-11-19
1                          How to Train Your Dragon      28.734
2010-03-26
2                                      Iron Man 2      28.515
2010-05-07
3                                       Toy Story      28.005
1995-11-22
4                                       Inception      27.920
2010-07-16
...                                           ...         ...
...
```

```
26512                              Laboratory Conditions       0.600
2018-10-13
26513                                   _EXHIBIT_84xxx_       0.600
2018-05-01
26514                                      The Last One       0.600
2018-10-01
26515                                      Trailer Made       0.600
2018-06-22
26516                                        The Church       0.600
2018-10-05

                                                   title  vote_average
vote_count
0      Harry Potter and the Deathly Hallows: Part 1           7.7
10788
1                           How to Train Your Dragon           7.7
7610
2                                        Iron Man 2           6.8
12368
3                                         Toy Story           7.9
10174
4                                         Inception           8.3
22186
...                                             ...           ...
...
26512                              Laboratory Conditions           0.0
1
26513                                   _EXHIBIT_84xxx_           0.0
1
26514                                      The Last One           0.0
1
26515                                      Trailer Made           0.0
1
26516                                        The Church           0.0
1

[26517 rows x 10 columns]
```

```
#Pandas is showing duplicate index columns so we drop one in the
following 2 columns
btmdb_df = btmdb_df.rename(columns={"unnamed:_0": "index"})

btmdb_df.iloc
```

```
<pandas.core.indexing._iLocIndexer at 0x212476e38b0>
```

```
btmdb_df = btmdb_df.set_index("index")

btmdb_df
```

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

                                      original_title  popularity
release_date  \
index

0      Harry Potter and the Deathly Hallows: Part 1      33.533
2010-11-19
1                          How to Train Your Dragon      28.734
2010-03-26
2                                       Iron Man 2      28.515
2010-05-07
3                                        Toy Story      28.005
1995-11-22
4                                        Inception      27.920
2010-07-16
...                                            ...         ...
...
26512                         Laboratory Conditions       0.600
2018-10-13
26513                                _EXHIBIT_84xxx_       0.600
2018-05-01
26514                                   The Last One       0.600
2018-10-01
26515                                   Trailer Made       0.600
2018-06-22
26516                                     The Church       0.600
2018-10-05

                                           title  vote_average
vote_count
index

0      Harry Potter and the Deathly Hallows: Part 1           7.7
10788
1                          How to Train Your Dragon           7.7
7610
2                                       Iron Man 2           6.8
```

```
12368
3                                        Toy Story           7.9
10174
4                                        Inception           8.3
22186
...                                            ...           ...
...
26512                          Laboratory Conditions           0.0
1
26513                                _EXHIBIT_84xxx_           0.0
1
26514                                   The Last One           0.0
1
26515                                   Trailer Made           0.0
1
26516                                     The Church           0.0
1

[26517 rows x 9 columns]
```

```
#looking at tmdb metadata
btmdb_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 26517 entries, 0 to 26516
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   genre_ids          26517 non-null  object 
 1   id                 26517 non-null  int64  
 2   original_language  26517 non-null  object 
 3   original_title     26517 non-null  object 
 4   popularity         26517 non-null  float64
 5   release_date       26517 non-null  object 
 6   title              26517 non-null  object 
 7   vote_average       26517 non-null  float64
 8   vote_count         26517 non-null  int64  
dtypes: float64(2), int64(2), object(5)
memory usage: 2.0+ MB
```

No need for cleaning

```
btmdb_df.describe()
```

```
                 id    popularity  vote_average    vote_count
count   26517.000000  26517.000000  26517.000000  26517.000000
mean   295050.153260      3.130912      5.991281    194.224837
std    153661.615648      4.355229      1.852946    960.961095
min        27.000000      0.600000      0.000000      1.000000
25%    157851.000000      0.600000      5.000000      2.000000
```

```
50%    309581.000000      1.374000      6.000000      5.000000
75%    419542.000000      3.694000      7.000000     28.000000
max    608444.000000     80.773000     10.000000  22186.000000

# Saving to cleaned database
btmdb_df.to_sql("tmdb_movies", cleaned_conn, if_exists="replace",
index=False)
```

5. Budgets

```
# Loading budget
bbudgets_df = pd.read_sql('''
    SELECT * FROM movie_budgets
''',conn)

bbudgets_df
```

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

      production_budget  domestic_gross  worldwide_gross
0             425000000.0     760507625.0     2.776345e+09
1             410600000.0     241063875.0     1.045664e+09
2             350000000.0      42762350.0     1.497624e+08
3             330600000.0     459005868.0     1.403014e+09
4             317000000.0     620181382.0     1.316722e+09
...                   ...             ...              ...
5777             7000.0             0.0     0.000000e+00
5778             6000.0         48482.0     2.404950e+05
5779             5000.0          1338.0     1.338000e+03
5780             1400.0             0.0     0.000000e+00
5781             1100.0        181041.0     1.810410e+05

[5782 rows x 6 columns]
```

```
#Checking for descriptive stats
bbudgets_df.describe()
```

```
                id  production_budget  domestic_gross  worldwide_gross
count  5782.000000       5.782000e+03    5.782000e+03     5.782000e+03
mean     50.372363       3.158776e+07    4.187333e+07     9.148746e+07
```

```
std       28.821076       4.181208e+07    6.824060e+07     1.747200e+08
min        1.000000       1.100000e+03    0.000000e+00     0.000000e+00
25%       25.000000       5.000000e+06    1.429534e+06     4.125415e+06
50%       50.000000       1.700000e+07    1.722594e+07     2.798445e+07
75%       75.000000       4.000000e+07    5.234866e+07     9.764584e+07
max      100.000000       4.250000e+08    9.366622e+08     2.776345e+09
```

3.Data analysis

Analyze ROI to identify which films and studios achieve the best returns

HO:there is no difference in average ROI between movie studios.

H1:there is a difference in average ROI between movie studios.

```
query_roi = """
SELECT
    mb.movie,
    bg.studio,
    ROUND((mb.worldwide_gross - mb.production_budget) /
mb.production_budget, 2) AS ROI,
    mb.worldwide_gross,
    mb.production_budget
FROM movie_budgets mb
JOIN bom_gross bg
    ON mb.movie = bg.title
WHERE mb.production_budget > 0
ORDER BY ROI DESC
LIMIT 10;
"""

broi_df = pd.read_sql(query_roi, cleaned_conn)
broi_df
```

```
                    movie   studio     ROI  worldwide_gross
production_budget
0             The Gallows  WB (NL)  415.56       41656474.0
100000.0
1       The Devil Inside     Par.  100.76      101759490.0
1000000.0
2              Insidious       FD   65.58       99870886.0
1500000.0
3             Unfriended     Uni.   63.36       64364198.0
1000000.0
4  Paranormal Activity 2     Par.   58.17      177512032.0
3000000.0
5                  Split     Uni.   54.79      278964806.0
5000000.0
6                Get Out     Uni.   50.07      255367951.0
```

```
5000000.0
7      Chernobyl Diaries       WB   41.41       42411721.0
1000000.0
8  Paranormal Activity 3     Par.   40.41      207039844.0
5000000.0
9              Annabelle  WB (NL)   38.52      256862920.0
6500000.0

broi_df.head()

                    movie   studio     ROI  worldwide_gross
production_budget
0            The Gallows  WB (NL)  415.56       41656474.0
100000.0
1       The Devil Inside     Par.  100.76      101759490.0
1000000.0
2              Insidious       FD   65.58       99870886.0
1500000.0
3             Unfriended     Uni.   63.36       64364198.0
1000000.0
4  Paranormal Activity 2     Par.   58.17      177512032.0
3000000.0

def hypothesis_test(ho,h1,p_value,alpha=0.05):
    if p_value < alpha:
        print(f"Reject the null hypothesis: {h0}")
        print(f"Accept the alternative hypothesis: {h1}")
    else:
        print(f"Fail to reject the null hypothesis: {h0}")
        print(f"Fail to accept the alternative hypothesis: {h1}")

from scipy.stats import f_oneway
# grouping movie studio by roi
groups=broi_df.groupby('studio')['ROI'].apply(list)
# run ANOVA
f_stat,p_value=f_oneway(*groups)
print(f"f_stat:",f_stat)
print(f"p_value:",p_value)

f_stat: 0.7790055085804414
p_value: 0.5838305729880893

# analyze ROI to identify which studios achieve the best returns
avg_roi_by_studio = broi_df.groupby('studio')
['ROI'].mean().sort_values(ascending=False)
avg_roi_by_studio

studio
WB (NL)    227.040000
Par.        66.446667
FD          65.580000
```

```
Uni.        56.073333
WB          41.410000
Name: ROI, dtype: float64
```

Finding

the WB(NL) studio has a very high profitability compaired to the others

```
h0="there is no difference in ROI between studios."
h1="there is a difference in ROI between studiods."
print(f"F-statistic={f_stat}, p-value={p_value}")
hypothesis_test(h0,h1,p_value)
```

```
F-statistic=0.7790055085804414, p-value=0.5838305729880893
Fail to reject the null hypothesis: there is no difference in ROI
between studios.
Fail to accept the alternative hypothesis: there is a difference in
ROI between studiods.
```

```
broi_df[broi_df['studio']=='WB (NL)']
```

```
         movie   studio     ROI  worldwide_gross  production_budget
0  The Gallows  WB (NL)  415.56       41656474.0           100000.0
9    Annabelle  WB (NL)   38.52      256862920.0          6500000.0
```

Recommendation

The studio should contain movies such as The gallows that fall under WB (NL) .

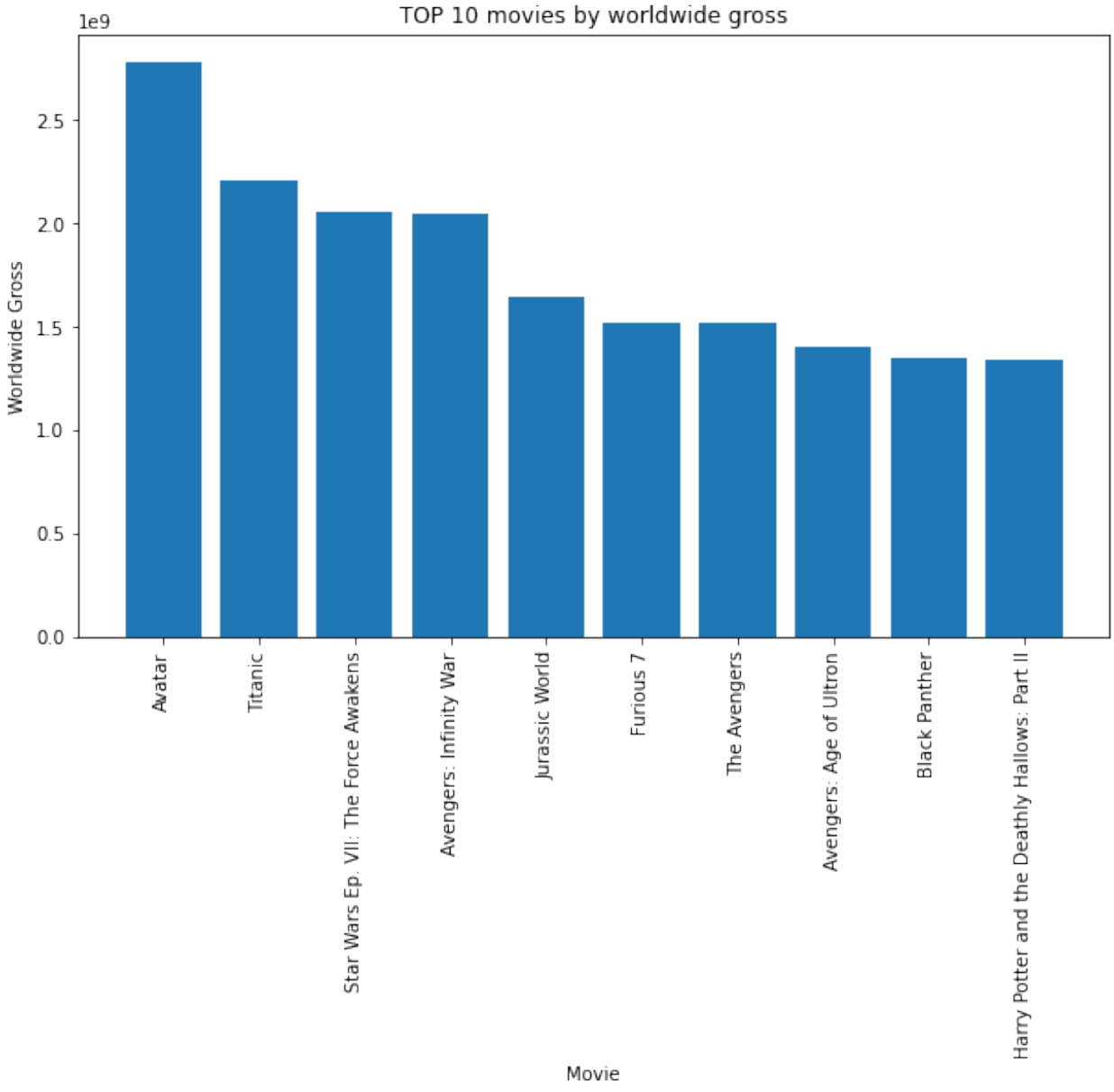
```
#Checking for null values
bbudgets_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5782 non-null   int64
 1   release_date       5782 non-null   object
 2   movie              5782 non-null   object
 3   production_budget  5782 non-null   float64
 4   domestic_gross     5782 non-null   float64
 5   worldwide_gross    5782 non-null   float64
dtypes: float64(3), int64(1), object(2)
memory usage: 271.2+ KB
```

```
TOP 10 movies by worldwide gross
```

```
top10 =bbudgets_df.nlargest(10,'worldwide_gross')
import matplotlib.pyplot as plt
```

```
import seaborn as sns
plt.figure(figsize=(10,6))
plt.bar(top10['movie'], top10['worldwide_gross'])
plt.xticks(rotation=90)
plt.title('TOP 10 movies by worldwide gross')
plt.xlabel('Movie')
plt.ylabel('Worldwide Gross')
plt.show()
```



Findings

Avatar movie has a high world wide gross than the other movies.

```
top10
```

```
     id  release_date                                                movie  \
0     1  Dec 18, 2009                                                Avatar
42   43  Dec 19, 1997                                               Titanic
5     6  Dec 18, 2015          Star Wars Ep. VII: The Force Awakens
6     7  Apr 27, 2018                      Avengers: Infinity War
33   34  Jun 12, 2015                                Jurassic World
66   67   Apr 3, 2015                                     Furious 7
26   27   May 4, 2012                                  The Avengers
3     4   May 1, 2015                      Avengers: Age of Ultron
41   42  Feb 16, 2018                                  Black Panther
260  61  Jul 15, 2011  Harry Potter and the Deathly Hallows: Part II


     production_budget  domestic_gross  worldwide_gross
0            425000000.0     760507625.0     2.776345e+09
42           200000000.0     659363944.0     2.208208e+09
5            306000000.0     936662225.0     2.053311e+09
6            300000000.0     678815482.0     2.048134e+09
33           215000000.0     652270625.0     1.648855e+09
66           190000000.0     353007020.0     1.518723e+09
26           225000000.0     623279547.0     1.517936e+09
3            330600000.0     459005868.0     1.403014e+09
41           200000000.0     700059566.0     1.348258e+09
260          125000000.0     381193157.0     1.341693e+09
```

Recomendation

For high chances of success the studio should consider high budget movies which tend to have a greater success.

To determine which genre is highly watched according to the runtime.

```
bmovieinfo_df.head()
```

```
   id                                           synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
```

```
4   7                                                  None     NR

                                  genre          director  \
0  Action and Adventure|Classics|Drama  William Friedkin
1    Drama|Science Fiction and Fantasy  David Cronenberg
2    Drama|Musical and Performing Arts    Allison Anders
3           Drama|Mystery and Suspense    Barry Levinson
4                        Drama|Romance    Rodney Bennett

                            writer  theater_date      dvd_date
box_office  \
0                  Ernest Tidyman   Oct 9, 1971  Sep 25, 2001
14141054.5
1     David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013
600000.0
2                   Allison Anders  Sep 13, 1996  Apr 18, 2000
14141054.5
3  Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997
14141054.5
4                     Giles Cooper          None          None
14141054.5

   runtime             studio  box_office_missing
0    104.0               None                True
1    108.0  Entertainment One               False
2    116.0               None                True
3    128.0               None                True
4    200.0               None                True
```

```
#to categorize runtime(short vs long)
def categorize_runtime(runtime):
    if runtime < 90:
        return 'Short'
    elif 90 <= runtime <= 150:
        return 'Medium'
    else:
        return 'Long'

#creating a new column indicating the runtime category per genre
bmovieinfo_df['runtime_category'] =
bmovieinfo_df['runtime'].apply(categorize_runtime)
bmovieinfo_df
```

```
        id                                           synopsis
rating  \
0        1  This gritty, fast-paced, and innovative police...      R

1        3  New York City, not-too-distant-future: Eric Pa...      R

2        5  Illeana Douglas delivers a superb performance ...      R
```

```
3        6  Michael Douglas runs afoul of a treacherous su...      R

4        7                                                None     NR

...    ...                                                 ...    ...

1555  1996  Forget terrorists or hijackers -- there's a ha...      R

1556  1997  The popular Saturday Night Live sketch was exp...     PG

1557  1998  Based on a novel by Richard Powell, when the l...      G

1558  1999  The Sandlot is a coming-of-age story about a g...     PG

1559  2000  Suspended from the force, Paris cop Hubert is ...      R


                                                    genre
director  \
0                    Action and Adventure|Classics|Drama    William
Friedkin
1                      Drama|Science Fiction and Fantasy    David
Cronenberg
2                      Drama|Musical and Performing Arts      Allison
Anders
3                             Drama|Mystery and Suspense      Barry
Levinson
4                                        Drama|Romance      Rodney
Bennett
...                                                  ...
...
1555   Action and Adventure|Horror|Mystery and Suspense
None
1556                  Comedy|Science Fiction and Fantasy        Steve
Barron
1557  Classics|Comedy|Drama|Musical and Performing Arts      Gordon
Douglas
1558    Comedy|Drama|Kids and Family|Sports and Fitness  David Mickey
Evans
1559  Action and Adventure|Art House and Internation...
None

                                                  writer
theater_date  \
0                                      Ernest Tidyman   Oct 9, 1971

1                         David Cronenberg|Don DeLillo  Aug 17, 2012

2                                      Allison Anders  Sep 13, 1996
```

```
3                      Paul Attanasio|Michael Crichton   Dec 9, 1994

4                                          Giles Cooper          None

...                                                 ...           ...

1555                                              None  Aug 18, 2006

1556  Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner  Jul 23, 1993

1557                                              None   Jan 1, 1962

1558                   David Mickey Evans|Robert Gunter   Apr 1, 1993

1559                                        Luc Besson  Sep 27, 2001


          dvd_date  box_office  runtime             studio  \
0     Sep 25, 2001  14141054.5    104.0               None
1      Jan 1, 2013    600000.0    108.0  Entertainment One
2     Apr 18, 2000  14141054.5    116.0               None
3     Aug 27, 1997  14141054.5    128.0               None
4             None  14141054.5    200.0               None
...            ...         ...      ...                ...
1555   Jan 2, 2007  33886034.0    106.0    New Line Cinema
1556  Apr 17, 2001  14141054.5     88.0  Paramount Vantage
1557  May 11, 2004  14141054.5    111.0               None
1558  Jan 29, 2002  14141054.5    101.0               None
1559  Feb 11, 2003  14141054.5     94.0  Columbia Pictures

      box_office_missing runtime_category
0                   True           Medium
1                  False           Medium
2                   True           Medium
3                   True           Medium
4                   True             Long
...                  ...              ...
1555               False           Medium
1556                True            Short
1557                True           Medium
1558                True           Medium
1559                True           Medium

[1560 rows x 13 columns]
```

```
#determine wich runtime groups are most watched per genre
runtime_genre_df = bmovieinfo_df.groupby(['genre',
'runtime_category']).size().reset_index(name='count')
runtime_genre_df.head()
```

```
                                                 genre runtime_category
count
0                                Action and Adventure           Medium
14
1                                Action and Adventure            Short
5
2  Action and Adventure|Animation|Art House and I...           Medium
1
3  Action and Adventure|Animation|Classics|Comedy...           Medium
1
4              Action and Adventure|Animation|Comedy           Medium
1
```

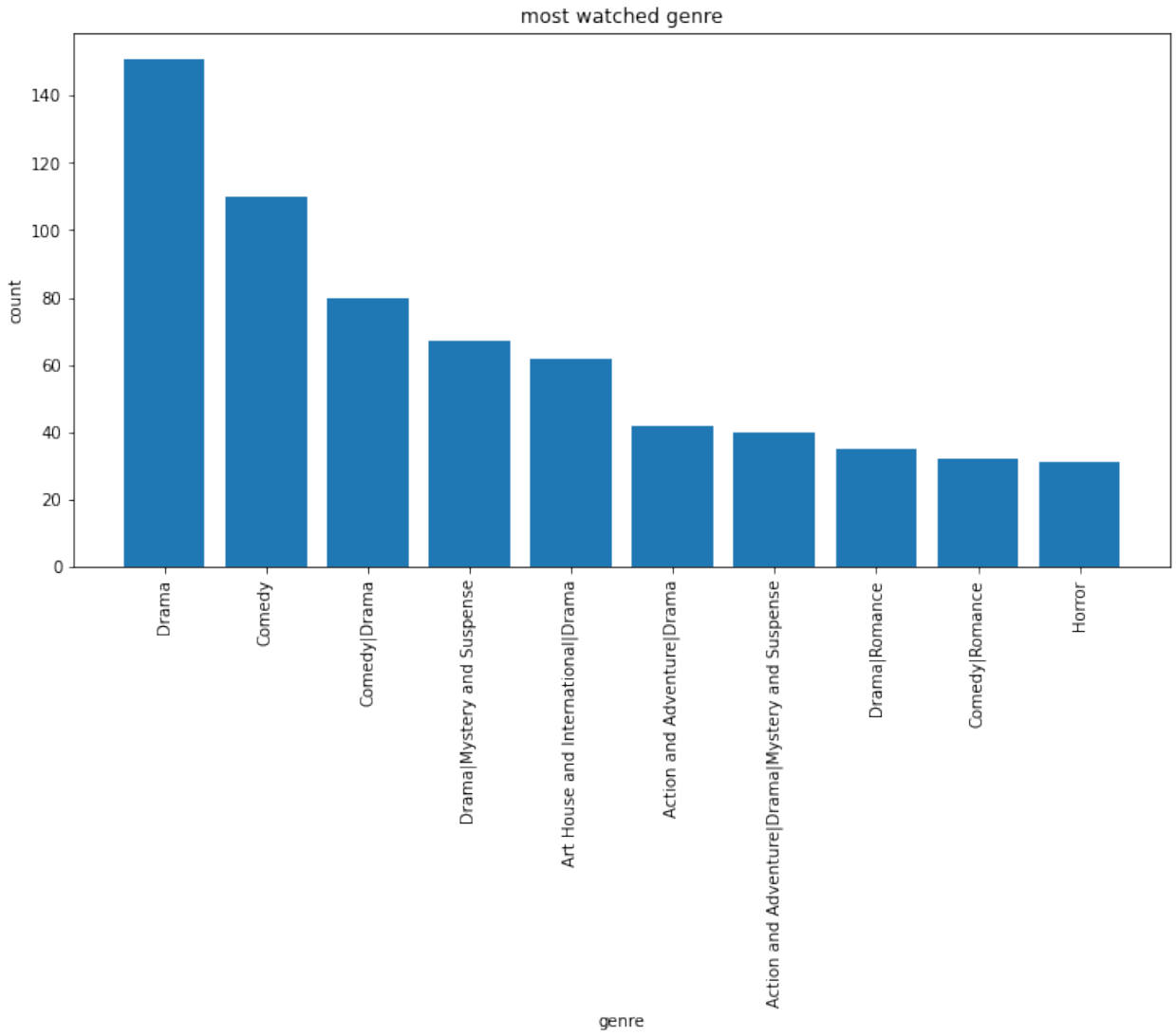
Findings

The action and adventure genre has the highest count compaired to the rest.meaning it is the most watched since the count represents the most watched genre per runtime category.

```
#bar graph:genre vs count
genre_counts=runtime_genre_df.groupby('genre')
['count'].sum().sort_values(ascending=False).head(10)
plt.figure(figsize=(12,6))
plt.bar(genre_counts.index,genre_counts.values)

plt.xlabel("genre")
plt.ylabel("count")
plt.title("most watched genre")
plt.xticks(rotation=90)

plt.show()
```



Recommendation

The the studio should contain this genres since they have the highest count per runtime .

Reginal revenue(market performance)

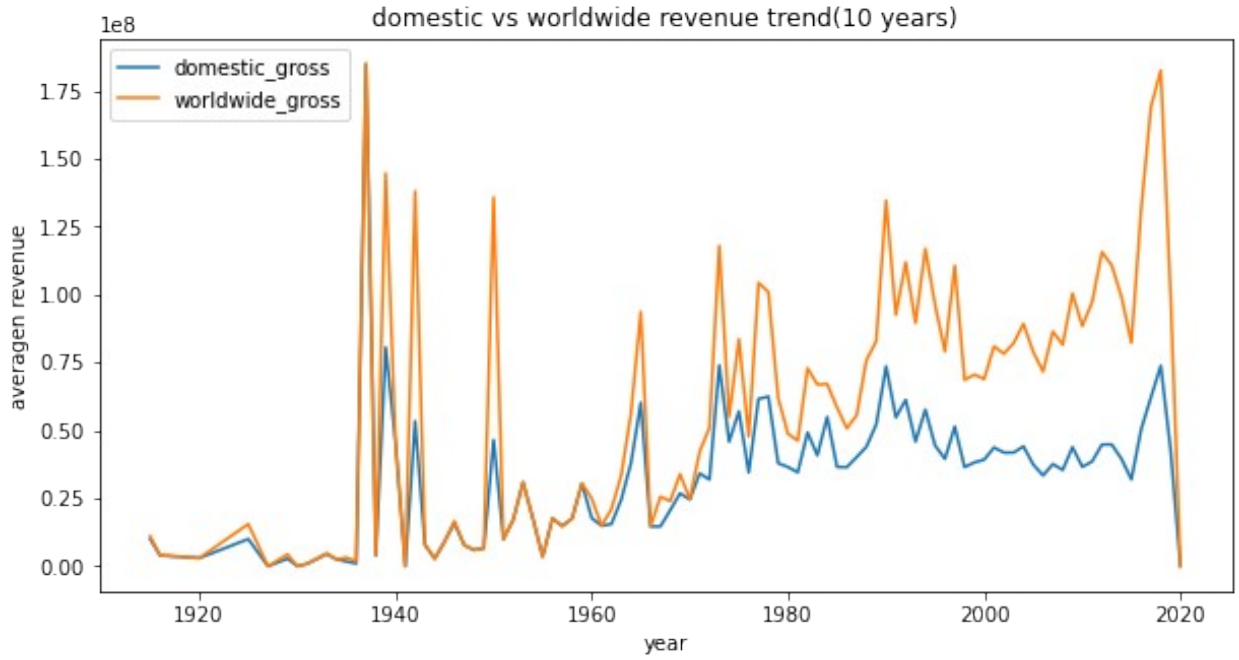
Revenue yearly trend (domestic vs international)

```
# extracting the years from the release dates.
bbudgets_df['release_date']=pd.to_datetime(bbudgets_df['release_date']
)
bbudgets_df['year']=bbudgets_df['release_date'].dt.year

#grouping the grosses according to the years
rev_region=bbudgets_df.groupby('year')
[['domestic_gross','worldwide_gross']].mean()
rev_region.head()
```

```
      domestic_gross  worldwide_gross
year
1915      10000000.0       11000000.0
1916       4000000.0        4000000.0
1920       3000000.0        3000000.0
1925      10000000.0       15500000.0
1927             0.0              0.0
```

```
rev_region.plot(figsize=(10,5))
plt.title('domestic vs worldwide revenue trend(10 years)')
plt.ylabel('averagen revenue')
plt.xlabel('year')
plt.show()
```



Findings

this shows that there is growth in the worldwide market compaired to the domestic market.

Recommendation

should consider world wide movies since they tend to have a greater income compaired to the others.

```
#Closing database connections
"""
WARNING! THIS SHOULD BE THE LAST CELL TO BE RAN SO AS TO AVOID ERRORS
"""
conn.close()
```

```
cleaned_conn.close()
im_conn.close()
```