Final Project Submission

Please fill out:

- Student name: AURALIA ADILLA MBOYA
- Student pace: Full Time
- Scheduled project review date/time: Nov 20th/ 11:59pm
- Instructor name: Mark Tiba
- Blog post URL: N/A

PROJECT OUTLINE

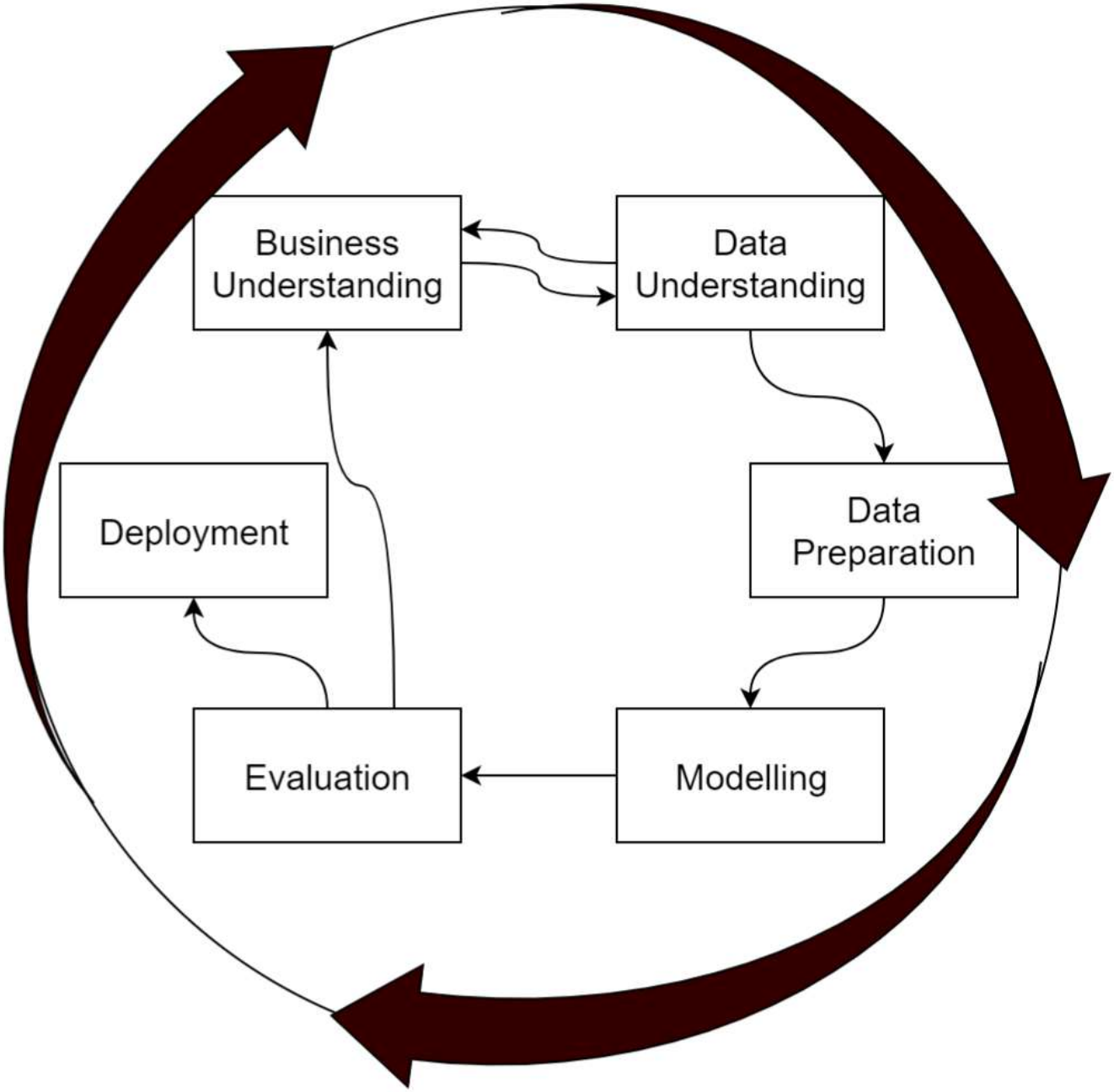
1. Introduction
2. Business Understanding
    - Business problem
    - Objectives
3. Data Understanding
4. Data preparation
    - Data Loading
    - Data cleaning
    - Data Analysis
5. Exploratory Descriptive Analysis (EDA)
    - Translating data into visual context
    - Plotting of graphs.
6. Conclusion
7. Recommendations

PHASE 1 PROJECT : Microsoft Film Prduction Studio

PROJECT OVERVIEW

I will use exploratory data analysis to produce insights for a business stakeholder in this segment.

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



BUSINESS UNDERSTANDING

Business Problem

Microsoft 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.They have hired you to help them better understand the movie industry. Your team is charged with exploring what type of films are currently doing the best at the box office. You must then translate those findings into actionable insights that the head of Microsoft's new movie studio can use to help decide what type of films to create.

Objectives

1. What is the correlation between the genre and movie runtime?
2. Which genre has the highest rating?
3. Which of the genres has the highest production budget?

4. Which Genre has the highest world wide gross?
5. Which is the most voted for genre?

DATA UNDERSTANDING

The majority of the information I used for this project came from a zipped folder that contained materials provided by the school. Since they have different file formats, they were all compressed into one folder.

The URLs to the data that I will be modifying for this project are listed below:

a. Box Office Mojo (https://www.boxofficemojo.com/)

b. IMDB (https://www.imdb.com/)

c. Rotten Tomatoes (https://www.rottentomatoes.com/)

d. TheMovieDB (https://www.themoviedb.org/)

e. The Numbers (https://www.the-numbers.com/)

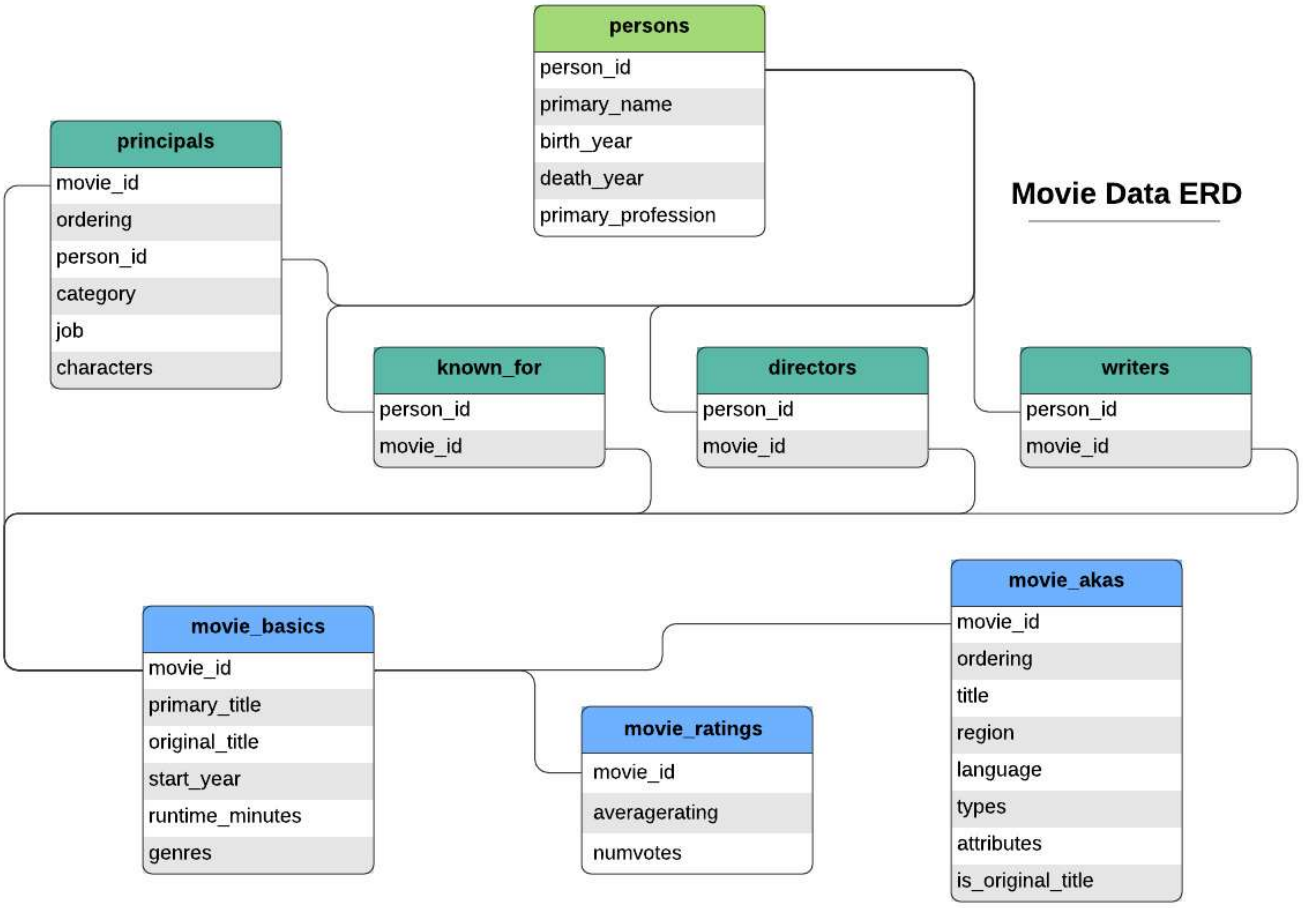
For a film studio to exist or be successful, we must conduct research, comprehend the information from the content provided, choose the right performers, and identify the top authors and writers for the various genres. To make Microsoft Film Studio successful, we will need to comprehend all the facts at our disposal. Four of These links' data were utilized for this project.

DATA PREPARATION

I'll be transforming data into usable format from this point on.

A TABLE OF RELATIONSHIPS

The relationships shown in the ERD below are what our datasets should have once they have been cleaned up in order for the stakeholder to understand what we are attempting to do.



LOADING DATA

In [1]:

```
# loading allthe necessary libraries
#For a more user-friendly data representation, import Pandas as pd.
#For the SQL database,import sqlite3
#import Numpy for arrays as np
#import Seaborn and Matplotlib for visualizations
#import json for the available structured data

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 [2]:

```
#Verifying that all necessary datasets have successfully loaded
#Checking for the necessary datasets
!ls -a
```

```
.
..
.canvas
.git
.gitignore
.ipynb_checkpoints
CONTRIBUTING.md
LICENSE.md
Production Budget  Vs Genres.png
Production Budget Vs Genres.png
README.md
awesome.gif
bom.movie_gross.csv
im.db
movie_data_erd.jpeg
rt.movie_info.tsv
rt.reviews.tsv
student.ipynb
tmdb.movies.csv
tn.movie_budgets.csv
untitled
zippedData
```

In [3]:

```
#loading the box office mojo file
movie_gross=pd.read_csv ('bom.movie_gross.csv')
movie_gross
```

Out[3]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000.0	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010
3	Inception	WB	292600000.0	535700000	2010
4	Shrek Forever After	P/DW	238700000.0	513900000	2010
...	...	...	...	...	...
3382	The Quake	Magn.	6200.0	NaN	2018
3383	Edward II (2018 re-release)	FM	4800.0	NaN	2018
3384	El Pacto	Sony	2500.0	NaN	2018
3385	The Swan	Synergetic	2400.0	NaN	2018
3386	An Actor Prepares	Grav.	1700.0	NaN	2018

3387 rows × 5 columns

In [4]:

```
#loading movie_budgets file
movie_budgets = pd.read_csv('tn.movie_budgets.csv')
movie_budgets
```

Out[4]:

	id	release_date	movie	production_budget	domestic_gross	worldwide_gross
0	1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
1	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
2	3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
3	4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
4	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747
...	...	...	...	...	...	...
5777	78	Dec 31, 2018	Red 11	$7,000	$0	$0
5778	79	Apr 2, 1999	Following	$6,000	$48,482	$240,495
5779	80	Jul 13, 2005	Return to the Land of Wonders	$5,000	$1,338	$1,338
5780	81	Sep 29, 2015	A Plague So Pleasant	$1,400	$0	$0
5781	82	Aug 5, 2005	My Date With Drew	$1,100	$181,041	$181,041

5782 rows × 6 columns

In [5]:

```
#loading the imdb file
movie_info= pd.read_csv('tmdb.movies.csv')
movie_info
```

Out[5]:

	Unnamed: 0	genre_ids	id	original_language	original_title	popularity	release_d
0	0	[12, 14, 10751]	12444	en	Harry Potter and the Deathly Hallows: Part 1	33.533	2010-11
1	1	[14, 12, 16, 10751]	10191	en	How to Train Your Dragon	28.734	2010-03
2	2	[12, 28, 878]	10138	en	Iron Man 2	28.515	2010-05
3	3	[16, 35, 10751]	862	en	Toy Story	28.005	1995-11
4	4	[28, 878, 12]	27205	en	Inception	27.920	2010-07
...	...	...	...	...	...	...	
26512	26512	[27, 18]	488143	en	Laboratory Conditions	0.600	2018-10
26513	26513	[18, 53]	485975	en	_EXHIBIT_84xxx_	0.600	2018-05
26514	26514	[14, 28, 12]	381231	en	The Last One	0.600	2018-10
26515	26515	[10751, 12, 28]	366854	en	Trailer Made	0.600	2018-06
26516	26516	[53, 27]	309885	en	The Church	0.600	2018-10

26517 rows × 10 columns

In [6]:

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

Out[6]:

	id	synopsis	rating	genre	director	writer	the
0	1	This gritty, fast-paced, and innovative police...	R	Action and Adventure|Classics|Drama	William Friedkin	Ernest Tidyman	C
1	3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	Au
2	5	Illeana Douglas delivers a superb performance ...	R	Drama|Musical and Performing Arts	Allison Anders	Allison Anders	Se
3	6	Michael Douglas runs afoul of a treacherous su...	R	Drama|Mystery and Suspense	Barry Levinson	Paul Attanasio|Michael Crichton	D
4	7	NaN	NR	Drama|Romance	Rodney Bennett	Giles Cooper	
...	...	...	...	...	...	...	
1555	1996	Forget terrorists or hijackers -- there's a ha...	R	Action and Adventure|Horror|Mystery and Suspense	NaN	NaN	Au
1556	1997	The popular Saturday Night Live sketch was exp...	PG	Comedy|Science Fiction and Fantasy	Steve Barron	Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner	Ju
1557	1998	Based on a novel by Richard Powell, when the l...	G	Classics|Comedy|Drama|Musical and Performing Arts	Gordon Douglas	NaN	Ja
1558	1999	The Sandlot is a coming-of-age story about a g...	PG	Comedy|Drama|Kids and Family|Sports and Fitness	David Mickey Evans	David Mickey Evans|Robert Gunter	A
1559	2000	Suspended from the force, Paris cop Hubert is ...	R	Action and Adventure|Art House and Internation...	NaN	Luc Besson	Se

1560 rows × 12 columns

In [7]:

```
#Using the encode attribute to load a tsv file and display tab separated values
rt_reviews = pd.read_table('rt.reviews.tsv', encoding='unicode_escape')
rt_reviews
```

Out[7]:

	id	review	rating	fresh	critic	top_critic	publisher	date
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018
1	3	It's an allegory in search of a meaning that n...	NaN	rotten	Annalee Newitz	0	io9.com	May 23, 2018
2	3	... life lived in a bubble in financial dealin...	NaN	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018
3	3	Continuing along a line introduced in last yea...	NaN	fresh	Daniel Kasman	0	MUBI	November 16, 2017
4	3	... a perverse twist on neorealism...	NaN	fresh	NaN	0	Cinema Scope	October 12, 2017
...	...	...	...	...	...	...	...	...
54427	2000	The real charm of this trifle is the deadpan c...	NaN	fresh	Laura Sinagra	1	Village Voice	September 24, 2002
54428	2000	NaN	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005
54429	2000	NaN	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005
54430	2000	NaN	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003
54431	2000	NaN	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002

54432 rows × 8 columns

In [8]:

```
#Sqlite3 connection to the database for reading the files
conn = sqlite3.connect("im.db")
conn
```

Out[8]:

```
<sqlite3.Connection at 0x2b2c795b3f0>
```

In [9]:

```
#load the necessary data from the movie_ratings sql file
movie_ratings = pd.read_sql_query("""
SELECT *
FROM movie_ratings
LIMIT 10
;
""", conn)
movie_ratings
```

Out[9]:

	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 [10]:

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

Out[10]:

	movie_id	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.0	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.0	Comedy,Drama,Fantasy
...	...	...	...	...	...	...
146139	tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.0	Drama
146140	tt9916622	Rodolpho Teóphilo - O Legado de um Pioneiro	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	NaN	Documentary
146141	tt9916706	Dankyavar Danka	Dankyavar Danka	2013	NaN	Comedy
146142	tt9916730	6 Gunn	6 Gunn	2017	116.0	None
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	NaN	Documentary

146144 rows × 6 columns

In [11]:

```
#load movie_akas to display relevant data needed.
movie_akas = pd.read_sql_query("""
SELECT *
FROM movie_akas
;
""", conn)
movie_akas
```

Out[11]:

	movie_id	ordering	title	region	language	types	attributes	is_original_ti
0	tt0369610	10	Джурасик свят	BG	bg	None	None	
1	tt0369610	11	Jurashikku warudo	JP	None	imdbDisplay	None	
2	tt0369610	12	Jurassic World: O Mundo dos Dinossauros	BR	None	imdbDisplay	None	
3	tt0369610	13	O Mundo dos Dinossauros	BR	None	None	short title	
4	tt0369610	14	Jurassic World	FR	None	imdbDisplay	None	
...	...	...	...	...	...	...	...	
331698	tt9827784	2	Sayonara kuchibiru	None	None	original	None	
331699	tt9827784	3	Farewell Song	XWW	en	imdbDisplay	None	
331700	tt9880178	1	La atención	None	None	original	None	
331701	tt9880178	2	La atención	ES	None	None	None	
331702	tt9880178	3	The Attention	XWW	en	imdbDisplay	None	

331703 rows × 8 columns

Data Cleaning

I'll be Correcting or deleting inaccurate, corrupted, improperly formatted, duplicate, or incomplete data from the relevant datasets.

In [12]:

```
#starting data cleaning from the first dataset
#cheecking for any erraneous data, null values or incomplete
movie_gross
```

Out[12]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000.0	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010
3	Inception	WB	292600000.0	535700000	2010
4	Shrek Forever After	P/DW	238700000.0	513900000	2010
...	...	...	...	...	...
3382	The Quake	Magn.	6200.0	NaN	2018
3383	Edward II (2018 re-release)	FM	4800.0	NaN	2018
3384	El Pacto	Sony	2500.0	NaN	2018
3385	The Swan	Synergetic	2400.0	NaN	2018
3386	An Actor Prepares	Grav.	1700.0	NaN	2018

3387 rows × 5 columns

In [13]:

```
#convert domestic_gross float to integer type
movie_gross['domestic_gross'] = movie_gross['domestic_gross'].fillna(0).astype(int)
```

In [14]:

```
#confirm the conversion
movie_gross
```

Out[14]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000	664300000	2010
3	Inception	WB	292600000	535700000	2010
4	Shrek Forever After	P/DW	238700000	513900000	2010
...	...	...	...	...	...
3382	The Quake	Magn.	6200	NaN	2018
3383	Edward II (2018 re-release)	FM	4800	NaN	2018
3384	El Pacto	Sony	2500	NaN	2018
3385	The Swan	Synergetic	2400	NaN	2018
3386	An Actor Prepares	Grav.	1700	NaN	2018

3387 rows × 5 columns

In [15]:

```
#convert domestic_gross float type to integer type
movie_gross['domestic_gross'].astype(int)
```

Out[15]:

```
0       415000000
1       334200000
2       296000000
3       292600000
4       238700000
          ...
3382         6200
3383         4800
3384         2500
3385         2400
3386         1700
Name: domestic_gross, Length: 3387, dtype: int32
```

In [16]:

```
#check for any null values
movie_gross.isna().sum()
```

Out[16]:

```
title                0
studio               5
domestic_gross       0
foreign_gross     1350
year                 0
dtype: int64
```

In [17]:

```
#checked for null values
#null values were found to be 1350 on foreign_gross column, 5 on studio
#Considering that they will be needed later, I have chosen to drop.
#drop all null values in the datasets
movie_gross.dropna()
```

Out[17]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415000000	652000000	2010
1	Alice in Wonderland (2010)	BV	334200000	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000	664300000	2010
3	Inception	WB	292600000	535700000	2010
4	Shrek Forever After	P/DW	238700000	513900000	2010
...	...	...	...	...	...
3275	I Still See You	LGF	1400	1500000	2018
3286	The Catcher Was a Spy	IFC	725000	229000	2018
3309	Time Freak	Grindstone	10000	256000	2018
3342	Reign of Judges: Title of Liberty - Concept Short	Darin Southa	93200	5200	2018
3353	Antonio Lopez 1970: Sex Fashion & Disco	FM	43200	30000	2018

2033 rows × 5 columns

In [18]:

```
#checking for any null values to clean
movie_budgets.isna().sum()
```

Out[18]:

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

In [19]:

```
#calling movie_budgets for cleaning
#checking for any null values
movie_budgets
```

Out[19]:

	id	release_date	movie	production_budget	domestic_gross	worldwide_gross
0	1	Dec 18, 2009	Avatar	$425,000,000	$760,507,625	$2,776,345,279
1	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	$410,600,000	$241,063,875	$1,045,663,875
2	3	Jun 7, 2019	Dark Phoenix	$350,000,000	$42,762,350	$149,762,350
3	4	May 1, 2015	Avengers: Age of Ultron	$330,600,000	$459,005,868	$1,403,013,963
4	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	$317,000,000	$620,181,382	$1,316,721,747
...	...	...	...	...	...	...
5777	78	Dec 31, 2018	Red 11	$7,000	$0	$0
5778	79	Apr 2, 1999	Following	$6,000	$48,482	$240,495
5779	80	Jul 13, 2005	Return to the Land of Wonders	$5,000	$1,338	$1,338
5780	81	Sep 29, 2015	A Plague So Pleasant	$1,400	$0	$0
5781	82	Aug 5, 2005	My Date With Drew	$1,100	$181,041	$181,041

5782 rows × 6 columns

In [20]:

```
#In the columns production budget, domestic gross, and worldwide gross,for movie_budgets da
movie_budgets['production_budget'] = movie_budgets['production_budget'].str.replace('$','')
movie_budgets['production_budget'] = movie_budgets['production_budget'].str.replace(',','')

movie_budgets['domestic_gross'] = movie_budgets['domestic_gross'].str.replace('$','')
movie_budgets['domestic_gross'] = movie_budgets['domestic_gross'].str.replace(',','')

movie_budgets['worldwide_gross'] = movie_budgets['worldwide_gross'].str.replace('$','')
movie_budgets['worldwide_gross'] = movie_budgets['worldwide_gross'].str.replace(',','')

movie_budgets
```

Out[20]:

	id	release_date	movie	production_budget	domestic_gross	worldwide_gross
0	1	Dec 18, 2009	Avatar	425000000	760507625	2776345279
1	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	410600000	241063875	1045663875
2	3	Jun 7, 2019	Dark Phoenix	350000000	42762350	149762350
3	4	May 1, 2015	Avengers: Age of Ultron	330600000	459005868	1403013963
4	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	317000000	620181382	1316721747
...	...	...	...	...	...	...
5777	78	Dec 31, 2018	Red 11	7000	0	0
5778	79	Apr 2, 1999	Following	6000	48482	240495
5779	80	Jul 13, 2005	Return to the Land of Wonders	5000	1338	1338
5780	81	Sep 29, 2015	A Plague So Pleasant	1400	0	0
5781	82	Aug 5, 2005	My Date With Drew	1100	181041	181041

5782 rows × 6 columns

In [21]:

```
#calling the next dataset, movie_info
movie_info
```

Out[21]:

	id	synopsis	rating	genre	director	writer	the
0	1	This gritty, fast-paced, and innovative police...	R	Action and Adventure|Classics|Drama	William Friedkin	Ernest Tidyman	C
1	3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	Au
2	5	Illeana Douglas delivers a superb performance ...	R	Drama|Musical and Performing Arts	Allison Anders	Allison Anders	Se
3	6	Michael Douglas runs afoul of a treacherous su...	R	Drama|Mystery and Suspense	Barry Levinson	Paul Attanasio|Michael Crichton	D
4	7	NaN	NR	Drama|Romance	Rodney Bennett	Giles Cooper	
...	...	...	...	...	...	...	
1555	1996	Forget terrorists or hijackers -- there's a ha...	R	Action and Adventure|Horror|Mystery and Suspense	NaN	NaN	Au
1556	1997	The popular Saturday Night Live sketch was exp...	PG	Comedy|Science Fiction and Fantasy	Steve Barron	Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner	Ju
1557	1998	Based on a novel by Richard Powell, when the l...	G	Classics|Comedy|Drama|Musical and Performing Arts	Gordon Douglas	NaN	Ja
1558	1999	The Sandlot is a coming-of-age story about a g...	PG	Comedy|Drama|Kids and Family|Sports and Fitness	David Mickey Evans	David Mickey Evans|Robert Gunter	A
1559	2000	Suspended from the force, Paris cop Hubert is ...	R	Action and Adventure|Art House and Internation...	NaN	Luc Besson	Se

1560 rows × 12 columns

In [22]:

```
#show all null values in the datasets
movie_info.isna().sum()
```

Out[22]:

```
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 [23]:

```
#The movie info dataset contains an excessive number of null values.
#synopsis 62,rating 3,genre8,director 199,writer 449,theater_date 359,dvd_date 359,currency

movie_info.dropna()
```

Out[23]:

	id	synopsis	rating	genre	director	writer	thea
1	3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	Aug
6	10	Some cast and crew from NBC's highly acclaimed...	PG-13	Comedy	Jake Kasdan	Mike White	Jar
7	13	Stewart Kane, an Irishman living in the Austra...	R	Drama	Ray Lawrence	Raymond Carver|Beatrix Christian	Apr
15	22	Two-time Academy Award Winner Kevin Spacey giv...	R	Comedy|Drama|Mystery and Suspense	George Hickenlooper	Norman Snider	Dec
18	25	From ancient Japan's most enduring tale, the e...	PG-13	Action and Adventure|Drama|Science Fiction and...	Carl Erik Rinsch	Chris Morgan|Hossein Amini	Dec
...	...	...	...	...	...	...	
1530	1968	This holiday season, acclaimed filmmaker Camer...	PG	Comedy|Drama	Cameron Crowe	Aline Brosh McKenna|Cameron Crowe	Dec
1537	1976	Embrace of the Serpent features the encounter,...	NR	Action and Adventure|Art House and International	Ciro Guerra	Ciro Guerra|Jacques Toulemonde Vidal	Feb
1541	1980	A band of renegades on the run in outer space ...	PG-13	Action and Adventure|Science Fiction and Fantasy	Joss Whedon	Joss Whedon	Sep
1542	1981	Money, Fame and the Knowledge of English. In I...	NR	Comedy|Drama	Gauri Shinde	Gauri Shinde	Oc

	id	synopsis	rating	genre	director	writer	thea
1545	1985	A woman who joins the undead against her will ...	R	Horror|Mystery and Suspense	Sebastian Gutierrez	Sebastian Gutierrez	Ju

235 rows × 12 columns

In [24]:

```
#cleaning data in rt_reviews
#call rt_reviews
rt_reviews
```

Out[24]:

	id	review	rating	fresh	critic	top_critic	publisher	date
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018
1	3	It's an allegory in search of a meaning that n...	NaN	rotten	Annalee Newitz	0	io9.com	May 23, 2018
2	3	... life lived in a bubble in financial dealin...	NaN	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018
3	3	Continuing along a line introduced in last yea...	NaN	fresh	Daniel Kasman	0	MUBI	November 16, 2017
4	3	... a perverse twist on neorealism...	NaN	fresh	NaN	0	Cinema Scope	October 12, 2017
...	...	...	...	...	...	...	...	...
54427	2000	The real charm of this trifle is the deadpan c...	NaN	fresh	Laura Sinagra	1	Village Voice	September 24, 2002
54428	2000	NaN	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005
54429	2000	NaN	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005
54430	2000	NaN	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003
54431	2000	NaN	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002

54432 rows × 8 columns

In [25]:

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

Out[25]:

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

In [26]:

```
#rt_reviews has too many null values
#review has 5563, rating 13517, critic 2722 and publisher has 309 null values
# drop all null values

rt_reviews.dropna()
```

Out[26]:

	id	review	rating	fresh	critic	top_critic	publisher	date
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018
6	3	Quickly grows repetitive and tiresome, meander...	C	rotten	Eric D. Snider	0	EricDSnider.com	July 17, 2013
7	3	Cronenberg is not a director to be daunted by ...	2/5	rotten	Matt Kelemen	0	Las Vegas CityLife	April 21, 2013
11	3	While not one of Cronenberg's stronger films, ...	B-	fresh	Emanuel Levy	0	EmanuelLevy.Com	February 3, 2013
12	3	Robert Pattinson works mighty hard to make Cos...	2/4	rotten	Christian Toto	0	Big Hollywood	January 15, 2013
...	...	...	...	...	...	...	...	...
54419	2000	Sleek, shallow, but frequently amusing.	2.5/4	fresh	Gene Seymour	1	Newsday	September 27, 2002
54420	2000	The spaniel-eyed Jean Reno infuses Hubert with...	3/4	fresh	Megan Turner	1	New York Post	September 27, 2002
54421	2000	Manages to be somewhat well-acted, not badly a...	1.5/4	rotten	Bob Strauss	0	Los Angeles Daily News	September 27, 2002
54422	2000	Arguably the best script that Besson has writt...	3.5/5	fresh	Wade Major	0	Boxoffice Magazine	September 27, 2002
54424	2000	Dawdles and drags when it should pop; it doesn...	1.5/5	rotten	Manohla Dargis	1	Los Angeles Times	September 26, 2002

33988 rows × 8 columns

Merging the relevant datasets

In [27]:

```
#checking for any NaN values
movie_ratings.isna().sum()
```

Out[27]:

```
movie_id         0
averagerating    0
numvotes         0
dtype: int64
```

In [28]:

```
#checking for any null values
movie_basics.isna().sum()
```

Out[28]:

```
movie_id              0
primary_title         0
original_title       21
start_year            0
runtime_minutes    31739
genres              5408
dtype: int64
```

In [29]:

```
#checking for null values
movie_akas.isna().sum()
```

Out[29]:

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

In [30]:

```
#eraaneous null values have been found in movie_akas
#replacing null values with 0
movie_akas.fillna(0, inplace= True)
movie_akas
```

Out[30]:

	movie_id	ordering	title	region	language	types	attributes	is_original_ti
0	tt0369610	10	Джурасик свят	BG	bg	0	0	
1	tt0369610	11	Jurashikku warudo	JP	0	imdbDisplay	0	
2	tt0369610	12	Jurassic World: O Mundo dos Dinossauros	BR	0	imdbDisplay	0	
3	tt0369610	13	O Mundo dos Dinossauros	BR	0	0	short title	
4	tt0369610	14	Jurassic World	FR	0	imdbDisplay	0	
...	...	...	...	...	...	...	...	
331698	tt9827784	2	Sayonara kuchibiru	0	0	original	0	
331699	tt9827784	3	Farewell Song	XWW	en	imdbDisplay	0	
331700	tt9880178	1	La atención	0	0	original	0	
331701	tt9880178	2	La atención	ES	0	0	0	
331702	tt9880178	3	The Attention	XWW	en	imdbDisplay	0	

331703 rows × 8 columns

In [31]:

```
#setting index of the dataframe
movie_ratings.set_index("movie_id")
```

Out[31]:

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 [32]:

```
#setting index for movies_basics
movie_basics.set_index("movie_id")
```

Out[32]:

movie_id	primary_title	original_title	start_year	runtime_minutes	genres
tt0063540	Sunghursh	Sunghursh	2013	175.0	Action,Crime,Drama
tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.0	Biography,Drama
tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.0	Drama
tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	NaN	Comedy,Drama
tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.0	Comedy,Drama,Fantasy
...	...	...	...	...	...
tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.0	Drama
tt9916622	Rodolpho Teóphilo - O Legado de um Pioneiro	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	NaN	Documentary
tt9916706	Dankyavar Danka	Dankyavar Danka	2013	NaN	Comedy
tt9916730	6 Gunn	6 Gunn	2017	116.0	None
tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	NaN	Documentary

146144 rows × 5 columns

In [33]:

```
#mergin movie_basics and movie_ratings
#call new table basics_and_ratings
basics_and_ratings = movie_ratings.merge(movie_basics, on = 'movie_id', how = 'inner')
basics_and_ratings
```

Out[33]:

	movie_id	averagerating	numvotes	primary_title	original_title	start_year	runtime_minutes
0	tt10356526	8.3	31	Laiye Je Yaarian	Laiye Je Yaarian	2019	117.0
1	tt10384606	8.9	559	Borderless	Borderless	2019	87.0
2	tt1042974	6.4	20	Just Inès	Just Inès	2010	90.0
3	tt1043726	4.2	50352	The Legend of Hercules	The Legend of Hercules	2014	99.0
4	tt1060240	6.5	21	Até Onde?	Até Onde?	2011	73.0
5	tt1069246	6.2	326	Habana Eva	Habana Eva	2010	106.0
6	tt1094666	7.0	1613	The Hammer	Hamill	2010	108.0
7	tt1130982	6.4	571	The Night Clerk	Avant l'aube	2011	104.0
8	tt1156528	7.2	265	Silent Sonata	Circus Fantasticus	2011	77.0
9	tt1161457	4.2	148	Vanquisher	The Vanquisher	2016	90.0

In [34]:

```
movie_akas.set_index('movie_id')
```

Out[34]:

movie_id	ordering	title	region	language	types	attributes	is_original_title
tt0369610	10	Джурасик свят	BG	bg	0	0	0.0
tt0369610	11	Jurashikku warudo	JP	0	imdbDisplay	0	0.0
tt0369610	12	Jurassic World: O Mundo dos Dinossauros	BR	0	imdbDisplay	0	0.0
tt0369610	13	O Mundo dos Dinossauros	BR	0	0	short title	0.0
tt0369610	14	Jurassic World	FR	0	imdbDisplay	0	0.0
...	...	...	...	...	...	...	...
tt9827784	2	Sayonara kuchibiru	0	0	original	0	1.0
tt9827784	3	Farewell Song	XWW	en	imdbDisplay	0	0.0
tt9880178	1	La atención	0	0	original	0	1.0
tt9880178	2	La atención	ES	0	0	0	0.0
tt9880178	3	The Attention	XWW	en	imdbDisplay	0	0.0

331703 rows × 7 columns

In [35]:

```
#merging basics_and_ratings & movie_akas
b_r_akas = basics_and_ratings.merge(movie_akas, on = 'movie_id', how= 'inner')
b_r_akas
```

Out[35]:

	movie_id	averagerating	numvotes	primary_title	original_title	start_year	runtime_minutes
0	tt1042974	6.4	20	Just Inès	Just Inès	2010	90.0
1	tt1042974	6.4	20	Just Inès	Just Inès	2010	90.0
2	tt1042974	6.4	20	Just Inès	Just Inès	2010	90.0
3	tt1043726	4.2	50352	The Legend of Hercules	The Legend of Hercules	2014	99.0
4	tt1043726	4.2	50352	The Legend of Hercules	The Legend of Hercules	2014	99.0
...	...	...	...	...	...	...	...
61	tt1156528	7.2	265	Silent Sonata	Circus Fantasticus	2011	77.0
62	tt1156528	7.2	265	Silent Sonata	Circus Fantasticus	2011	77.0
63	tt1156528	7.2	265	Silent Sonata	Circus Fantasticus	2011	77.0
64	tt1161457	4.2	148	Vanquisher	The Vanquisher	2016	90.0
65	tt1161457	4.2	148	Vanquisher	The Vanquisher	2016	90.0

66 rows × 15 columns

In [36]:

```
#setting index for movie_budgets
movie_budgets.set_index('domestic_gross','production_budget')
```

Out[36]:

domestic_gross	id	release_date	movie	production_budget	worldwide_gross
760507625	1	Dec 18, 2009	Avatar	425000000	2776345279
241063875	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides	410600000	1045663875
42762350	3	Jun 7, 2019	Dark Phoenix	350000000	149762350
459005868	4	May 1, 2015	Avengers: Age of Ultron	330600000	1403013963
620181382	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi	317000000	1316721747
...	...	...	...	...	...
0	78	Dec 31, 2018	Red 11	7000	0
48482	79	Apr 2, 1999	Following	6000	240495
1338	80	Jul 13, 2005	Return to the Land of Wonders	5000	1338
0	81	Sep 29, 2015	A Plague So Pleasant	1400	0
181041	82	Aug 5, 2005	My Date With Drew	1100	181041

5782 rows × 5 columns

In [37]:

```
#setting index for movie_gross
movie_gross.set_index('domestic_gross', 'production_budget')
```

Out[37]:

domestic_gross	title	studio	foreign_gross	year
415000000	Toy Story 3	BV	652000000	2010
334200000	Alice in Wonderland (2010)	BV	691300000	2010
296000000	Harry Potter and the Deathly Hallows Part 1	WB	664300000	2010
292600000	Inception	WB	535700000	2010
238700000	Shrek Forever After	P/DW	513900000	2010
...	...	...	...	...
6200	The Quake	Magn.	NaN	2018
4800	Edward II (2018 re-release)	FM	NaN	2018
2500	El Pacto	Sony	NaN	2018
2400	The Swan	Synergetic	NaN	2018
1700	An Actor Prepares	Grav.	NaN	2018

3387 rows × 4 columns

In [38]:

```
#merging tables to access data based on the logical relationships between them
#merging the movie_basics and movie_ratings
#call new table ratings_basics
joined_gross_budget = pd.concat([movie_gross,movie_budgets], axis=1)
joined_gross_budget
```

Out[38]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	NaN	NaN	NaN	NaN	NaN	78	Dec 31, 2018	Red 11
5778	NaN	NaN	NaN	NaN	NaN	79	Apr 2, 1999	Following
5779	NaN	NaN	NaN	NaN	NaN	80	Jul 13, 2005	Return to the Land of Wonders
5780	NaN	NaN	NaN	NaN	NaN	81	Sep 29, 2015	A Plague So Pleasant
5781	NaN	NaN	NaN	NaN	NaN	82	Aug 5, 2005	My Date With Drew

5782 rows × 11 columns

In [39]:

```
#merging joined_gross_budget,b_r_akas
akas_gross = pd.concat([joined_gross_budget,b_r_akas], axis=1)
akas_gross
```

Out[39]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	NaN	NaN	NaN	NaN	NaN	78	Dec 31, 2018	Red 11
5778	NaN	NaN	NaN	NaN	NaN	79	Apr 2, 1999	Following
5779	NaN	NaN	NaN	NaN	NaN	80	Jul 13, 2005	Return to the Land of Wonders
5780	NaN	NaN	NaN	NaN	NaN	81	Sep 29, 2015	A Plague So Pleasant
5781	NaN	NaN	NaN	NaN	NaN	82	Aug 5, 2005	My Date With Drew

5782 rows × 26 columns

In [40]:

```
#setting index for rt_reviews dataframe
rt_reviews.set_index('id')
```

Out[40]:

id	review	rating	fresh	critic	top_critic	publisher	date
3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018
3	It's an allegory in search of a meaning that n...	NaN	rotten	Annalee Newitz	0	io9.com	May 23, 2018
3	... life lived in a bubble in financial dealin...	NaN	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018
3	Continuing along a line introduced in last yea...	NaN	fresh	Daniel Kasman	0	MUBI	November 16, 2017
3	... a perverse twist on neorealism...	NaN	fresh	NaN	0	Cinema Scope	October 12, 2017
...	...	...	...	...	...	...	...
2000	The real charm of this trifle is the deadpan c...	NaN	fresh	Laura Sinagra	1	Village Voice	September 24, 2002
2000	NaN	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005
2000	NaN	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005
2000	NaN	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003
2000	NaN	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002

54432 rows × 7 columns

In [41]:

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

Out[41]:

id	synopsis	rating	genre	director	writer	theater_d
1	This gritty, fast-paced, and innovative police...	R	Action and Adventure|Classics|Drama	William Friedkin	Ernest Tidyman	Oct 9, 19
3	New York City, not-too-distant-future: Eric Pa...	R	Drama|Science Fiction and Fantasy	David Cronenberg	David Cronenberg|Don DeLillo	Aug 17, 20
5	Illeana Douglas delivers a superb performance ...	R	Drama|Musical and Performing Arts	Allison Anders	Allison Anders	Sep 13, 19
6	Michael Douglas runs afoul of a treacherous su...	R	Drama|Mystery and Suspense	Barry Levinson	Paul Attanasio|Michael Crichton	Dec 9, 19
7	NaN	NR	Drama|Romance	Rodney Bennett	Giles Cooper	N
...	...	...	...	...	...	
1996	Forget terrorists or hijackers -- there's a ha...	R	Action and Adventure|Horror|Mystery and Suspense	NaN	NaN	Aug 18, 20
1997	The popular Saturday Night Live sketch was exp...	PG	Comedy|Science Fiction and Fantasy	Steve Barron	Terry Turner|Tom Davis|Dan Aykroyd|Bonnie Turner	Jul 23, 19
1998	Based on a novel by Richard Powell, when the l...	G	Classics|Comedy|Drama|Musical and Performing Arts	Gordon Douglas	NaN	Jan 1, 19
1999	The Sandlot is a coming-of-age story about a g...	PG	Comedy|Drama|Kids and Family|Sports and Fitness	David Mickey Evans	David Mickey Evans|Robert Gunter	Apr 1, 19
2000	Suspended from the force, Paris cop Hubert is ...	R	Action and Adventure|Art House and Internation...	NaN	Luc Besson	Sep 27, 20

1560 rows × 11 columns

Using .dropna() in the merged datasets. The dataframes have crossed the threshold of null values, thus dropping.

In [42]:

```
#merging rt_reviews and movie_info datasets
reviews_info = pd.concat([rt_reviews,movie_info], axis=1)
reviews_info
```

Out[42]:

	id	review	rating	fresh	critic	top_critic	publisher	date	
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018	
1	3	It's an allegory in search of a meaning that n...	NaN	rotten	Annalee Newitz	0	io9.com	May 23, 2018	
2	3	... life lived in a bubble in financial dealin...	NaN	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018	
3	3	Continuing along a line introduced in last yea...	NaN	fresh	Daniel Kasman	0	MUBI	November 16, 2017	
4	3	... a perverse twist on neorealism...	NaN	fresh	NaN	0	Cinema Scope	October 12, 2017	
...	...	...	...	...	...	...	...	...	
54427	2000	The real charm of this trifle is the deadpan c...	NaN	fresh	Laura Sinagra	1	Village Voice	September 24, 2002	N
54428	2000	NaN	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005	N
54429	2000	NaN	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005	N
54430	2000	NaN	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003	N
54431	2000	NaN	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002	N

54432 rows × 20 columns

In [43]:

```
#merged tthe two dataframes
#drop all the NaN values a
reviews_info.dropna()
```

Out[43]:

	id	review	rating	fresh	critic	top_critic	publisher	date
6	3	Quickly grows repetitive and tiresome, meander...	C	rotten	Eric D. Snider	0	EricDSnider.com	July 17, 2013
7	3	Cronenberg is not a director to be daunted by ...	2/5	rotten	Matt Kelemen	0	Las Vegas CityLife	April 21, 2013
15	3	For better or worse - often both - Cosmopolis ...	3/5	fresh	Adam Ross	0	The Aristocrat	September 27, 2012
18	3	It's fascinating to watch Pattinson actually a...	2/4	rotten	Sean P. Means	0	Salt Lake Tribune	September 14, 2012
19	3	A black comedy as dry and deadpan as a bleache...	4/4	fresh	John Beifuss	0	Commercial Appeal (Memphis, TN)	September 10, 2012
...	...	...	...	...	...	...	...	...
1511	45	Hello, Deedles. Terrible to meet you.	1/5	rotten	Scott Weinberg	0	eFilmCritic.com	July 29, 2002
1518	45	Steve Van Wormer and Paul Walker, as Stew and ...	0/4	rotten	Steve Rhodes	0	Internet Reviews	January 1, 2000
1537	46	Leaves the audience smiling and giggling, all ...	3/4	fresh	Michael Dequina	0	TheMovieReport.com	March 8, 2009
1541	46	The briskly paced, high-spirited movie is comp...	3.5/4	fresh	Judith Egerton	0	Courier-Journal (Louisville, KY)	June 25, 2004

	id	review	rating	fresh	critic	top_critic	publisher	date
1545	46	It's a familiar show-biz routine but one that'...	3.5/4	fresh	Susan Wloszczyna	1	USA Today	January 1, 2000

148 rows × 20 columns

In [44]:

```
#merge the joined_gross_budget with basics_and_ratings
budget_ratings = pd.concat([joined_gross_budget,basics_and_ratings], axis=1)
budget_ratings
```

Out[44]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	NaN	NaN	NaN	NaN	NaN	78	Dec 31, 2018	Red 11
5778	NaN	NaN	NaN	NaN	NaN	79	Apr 2, 1999	Following
5779	NaN	NaN	NaN	NaN	NaN	80	Jul 13, 2005	Return to the Land of Wonders
5780	NaN	NaN	NaN	NaN	NaN	81	Sep 29, 2015	A Plague So Pleasant
5781	NaN	NaN	NaN	NaN	NaN	82	Aug 5, 2005	My Date With Drew

5782 rows × 19 columns

In [45]:

```
#drop the null values.
budget_ratings.fillna(0, inplace = True)
budget_ratings
```

Out[45]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	0	0	0.0	0	0.0	78	Dec 31, 2018	Red 11
5778	0	0	0.0	0	0.0	79	Apr 2, 1999	Following
5779	0	0	0.0	0	0.0	80	Jul 13, 2005	Return to the Land of Wonders
5780	0	0	0.0	0	0.0	81	Sep 29, 2015	A Plague So Pleasant
5781	0	0	0.0	0	0.0	82	Aug 5, 2005	My Date With Drew

5782 rows × 19 columns

In [46]:

```
#merging all the dataframes
#merging akas_gross, budgets_ratings
budget_ratings_akas = pd.concat([reviews_info,budget_ratings], axis=1)
budget_ratings_akas
```

Out[46]:

	id	review	rating	fresh	critic	top_critic	publisher	date	
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018	
1	3	It's an allegory in search of a meaning that n...	NaN	rotten	Annalee Newitz	0	io9.com	May 23, 2018	
2	3	... life lived in a bubble in financial dealin...	NaN	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018	
3	3	Continuing along a line introduced in last yea...	NaN	fresh	Daniel Kasman	0	MUBI	November 16, 2017	
4	3	... a perverse twist on neorealism...	NaN	fresh	NaN	0	Cinema Scope	October 12, 2017	
...	...	...	...	...	...	...	...	...	
54427	2000	The real charm of this trifle is the deadpan c...	NaN	fresh	Laura Sinagra	1	Village Voice	September 24, 2002	N
54428	2000	NaN	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005	N
54429	2000	NaN	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005	N
54430	2000	NaN	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003	N
54431	2000	NaN	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002	N

54432 rows × 39 columns

In [47]:

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

Out[47]:

	id	review	rating	fresh	critic	top_critic	publisher	date	
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018	1
1	3	It's an allegory in search of a meaning that n...	0	rotten	Annalee Newitz	0	io9.com	May 23, 2018	3
2	3	... life lived in a bubble in financial dealin...	0	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018	5
3	3	Continuing along a line introduced in last yea...	0	fresh	Daniel Kasman	0	MUBI	November 16, 2017	6
4	3	... a perverse twist on neorealism...	0	fresh	0	0	Cinema Scope	October 12, 2017	7
...	...	...	...	...	...	...	...	...	
54427	2000	The real charm of this trifle is the deadpan c...	0	fresh	Laura Sinagra	1	Village Voice	September 24, 2002	0
54428	2000	0	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005	0
54429	2000	0	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005	0
54430	2000	0	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003	0
54431	2000	0	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002	0

54432 rows × 39 columns

Exploratory Descriptive Analysis (EDA)

We'll now employ techniques that are sometimes referred to as descriptive statistics because they only describe the available data or offer estimations based on it.

What is the correlation between the genre and movie runtime?

In [48]:

```
#open the needed dataframe
budget_ratings_akas
```

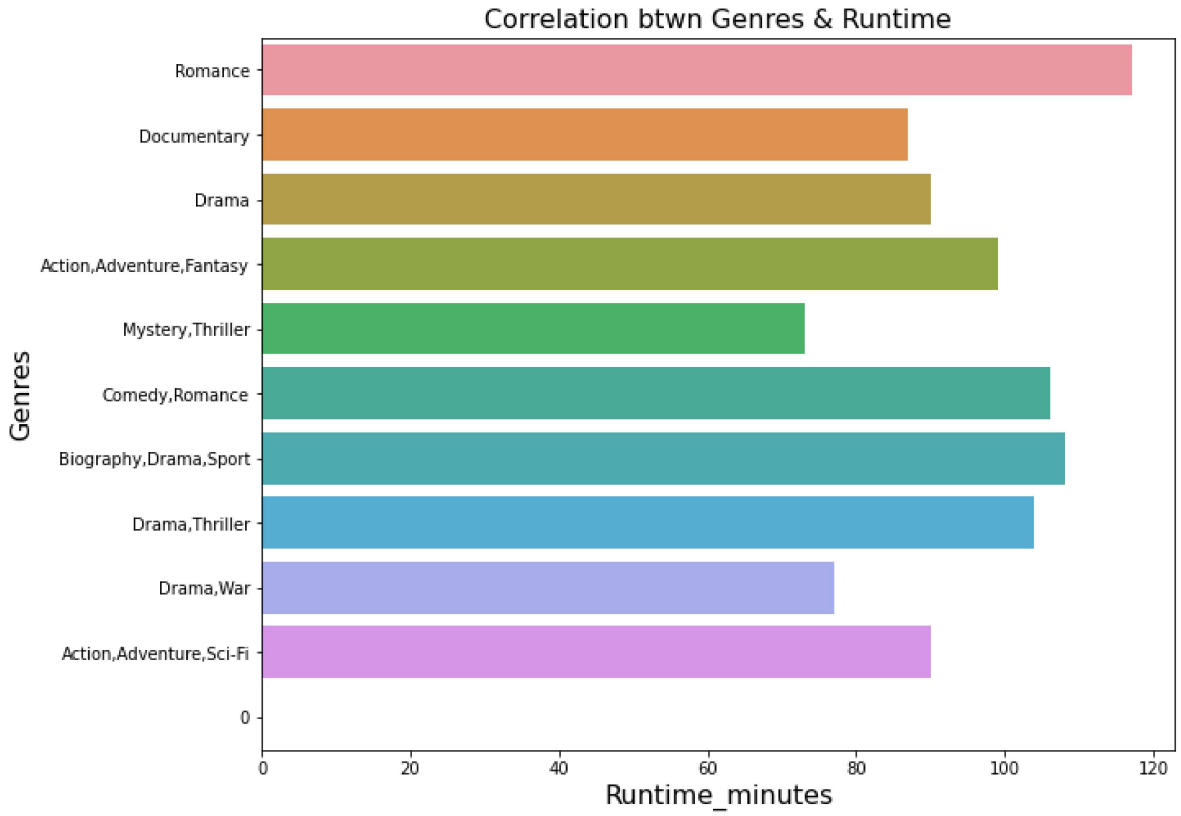
Out[48]:

	id	review	rating	fresh	critic	top_critic	publisher	date	
0	3	A distinctly gallows take on contemporary fina...	3/5	fresh	PJ Nabarro	0	Patrick Nabarro	November 10, 2018	1
1	3	It's an allegory in search of a meaning that n...	0	rotten	Annalee Newitz	0	io9.com	May 23, 2018	3
2	3	... life lived in a bubble in financial dealin...	0	fresh	Sean Axmaker	0	Stream on Demand	January 4, 2018	5
3	3	Continuing along a line introduced in last yea...	0	fresh	Daniel Kasman	0	MUBI	November 16, 2017	6
4	3	... a perverse twist on neorealism...	0	fresh	0	0	Cinema Scope	October 12, 2017	7
...	...	...	...	...	...	...	...	...	
54427	2000	The real charm of this trifle is the deadpan c...	0	fresh	Laura Sinagra	1	Village Voice	September 24, 2002	0
54428	2000	0	1/5	rotten	Michael Szymanski	0	Zap2it.com	September 21, 2005	0
54429	2000	0	2/5	rotten	Emanuel Levy	0	EmanuelLevy.Com	July 17, 2005	0
54430	2000	0	2.5/5	rotten	Christopher Null	0	Filmcritic.com	September 7, 2003	0
54431	2000	0	3/5	fresh	Nicolas Lacroix	0	Showbizz.net	November 12, 2002	0

54432 rows × 39 columns

In [49]:

```
# plotting a sns.barplot:
fig, ax1= plt.subplots(figsize=(10,8))

x = list(budget_ratings_akas['runtime_minutes'].values)
y = budget_ratings['genres']

ax= sns.barplot(data = budget_ratings, x = 'runtime_minutes', y = 'genres')

#labelling plot
ax1.set_title('Correlation btwn Genres & Runtime', fontsize=16)
ax1.set_xlabel("Runtime_minutes",fontsize=16)
ax1.set_ylabel("Genres", fontsize=16)

#will display the plot
plt.show()
```



The plot above has the longest duration of the genres, measured in minutes, according to the visual representation. The genre with the longest runtime is romance, whereas the genre with the shortest length is a thriller.

Which genre has the highest rating?

In [50]:

```
#loading data
#confirming the columns needed are available
budget_ratings
```

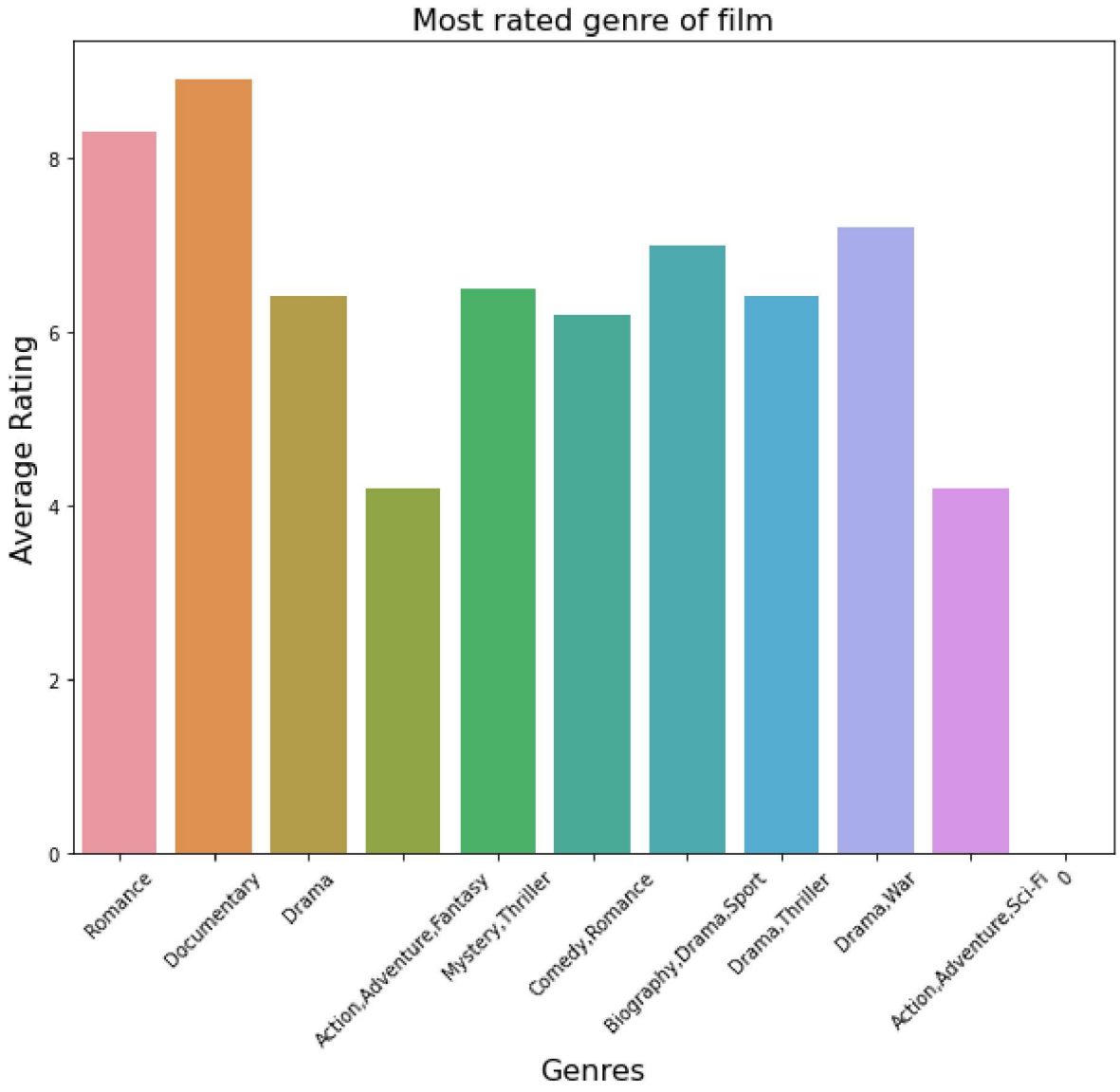
Out[50]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	0	0	0.0	0	0.0	78	Dec 31, 2018	Red 11
5778	0	0	0.0	0	0.0	79	Apr 2, 1999	Following
5779	0	0	0.0	0	0.0	80	Jul 13, 2005	Return to the Land of Wonders
5780	0	0	0.0	0	0.0	81	Sep 29, 2015	A Plague So Pleasant
5781	0	0	0.0	0	0.0	82	Aug 5, 2005	My Date With Drew

5782 rows × 19 columns

In [51]:

```
#plotting
fig, ax1= plt.subplots(figsize=(10,8))

x = list(budget_ratings_akas['genres'].values)
y = budget_ratings['averagerating']

ax= sns.barplot(data = budget_ratings, x = 'genres', y = 'averagerating')

#labelling plot
ax1.set_title('Most rated genre of film', fontsize=16)
ax1.set_xlabel("Genres",fontsize=16)
ax1.set_ylabel("Average Rating", fontsize=16)

#changing axis of x labels
plt.xticks(rotation = 45)

#will display the plot
plt.show()
```



The genre with the highest rating, documentaries, is 8.9

Which of the genres has the highest world wide gross?

In [52]:

```
#To enable it to load in the plot, convert worldwide gross to float.
budget_ratings['worldwide_gross']=budget_ratings['worldwide_gross'].astype(float)
```

In [53]:

```
#confirming changes
budget_ratings
```

Out[53]:

	title	studio	domestic_gross	foreign_gross	year	id	release_date	movie
0	Toy Story 3	BV	415000000.0	652000000	2010.0	1	Dec 18, 2009	Avatar
1	Alice in Wonderland (2010)	BV	334200000.0	691300000	2010.0	2	May 20, 2011	Pirates of the Caribbean: On Stranger Tides
2	Harry Potter and the Deathly Hallows Part 1	WB	296000000.0	664300000	2010.0	3	Jun 7, 2019	Dark Phoenix
3	Inception	WB	292600000.0	535700000	2010.0	4	May 1, 2015	Avengers: Age of Ultron
4	Shrek Forever After	P/DW	238700000.0	513900000	2010.0	5	Dec 15, 2017	Star Wars Ep. VIII: The Last Jedi
...	...	...	...	...	...	...	...	...
5777	0	0	0.0	0	0.0	78	Dec 31, 2018	Red 11
5778	0	0	0.0	0	0.0	79	Apr 2, 1999	Following
5779	0	0	0.0	0	0.0	80	Jul 13, 2005	Return to the Land of Wonders
5780	0	0	0.0	0	0.0	81	Sep 29, 2015	A Plague So Pleasant
5781	0	0	0.0	0	0.0	82	Aug 5, 2005	My Date With Drew

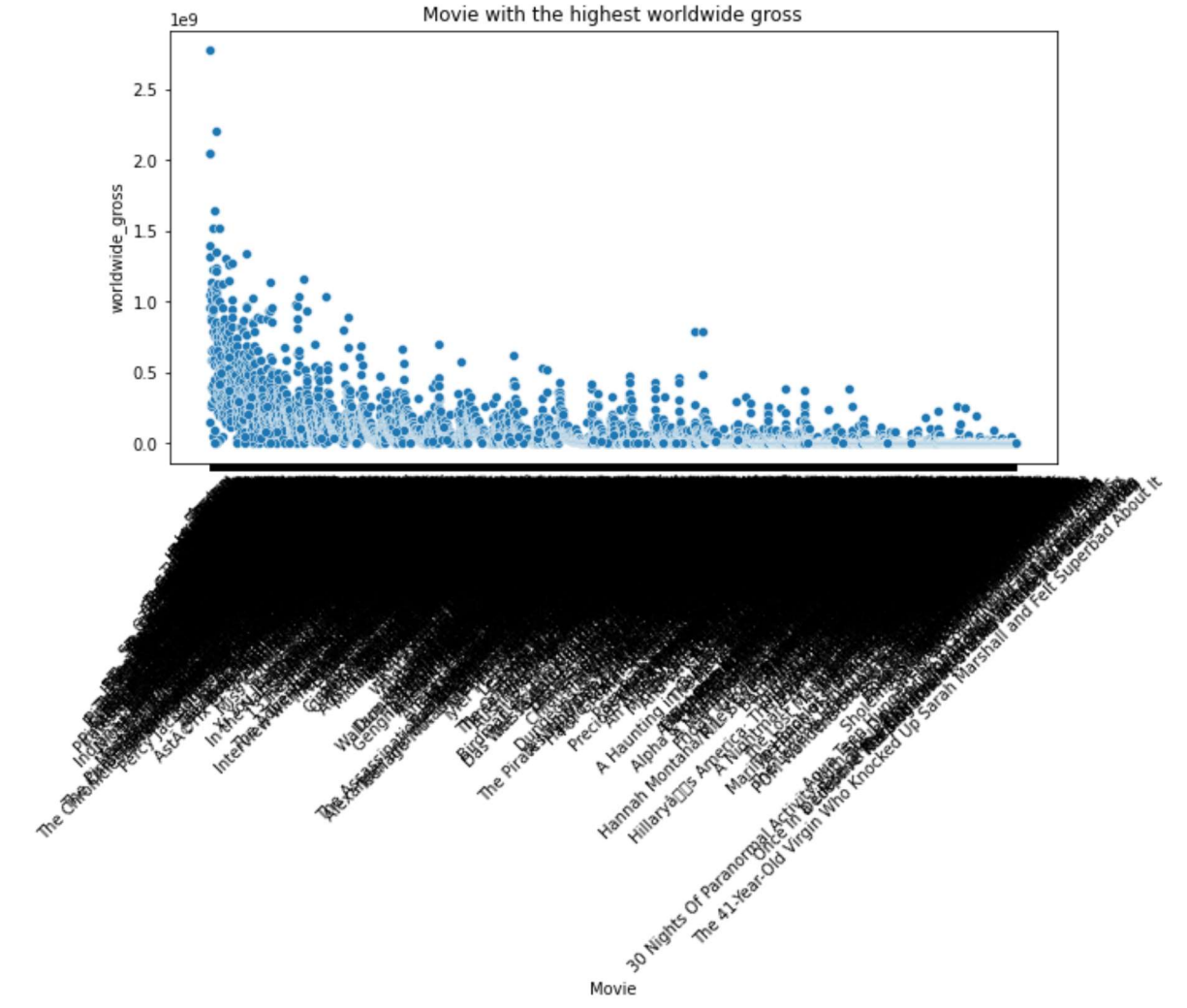
5782 rows × 19 columns

In [54]:

```
fig, ax1= plt.subplots(figsize=(10,5))

#arranging the x & y axis to avoid an overlap
x = np.arange(8)
y = 2*x + 1


#plot:
ax= sns.scatterplot( x='movie', y='worldwide_gross', data = budget_ratings)


#labelling plot
ax1.set_title('Movie with the highest worldwide gross')
ax1.set_xlabel("Movie")
ax1.set_ylabel("worldwide_gross")
plt.xticks(rotation= 45)
#will display the plot
plt.show()
```

```
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 128 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 153 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 148 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 129 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 149 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 159 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 131 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:238: RuntimeWarning: Glyph 147 missing from current f
ont.
  font.set_text(s, 0.0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 128 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 153 missing from current f
ont.
```

```
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 148 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 129 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 149 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 159 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 131 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
C:\Users\AURALIA\anaconda3\envs\learn-env\lib\site-packages\matplotlib\bac
kends\backend_agg.py:201: RuntimeWarning: Glyph 147 missing from current f
ont.
  font.set_text(s, 0, flags=flags)
```



This particular scatter plot was created to display the amount of money that the films on the x-axis brought in globally. I made a few attempts to stop it from overlapping, but they were unsuccessful. This leads me to the conclusion that I need to do additional research on how to plan a plot that doesn't overlap. The many film

genres brought in good money as gained from x-axis, which is measured in millions.

Which movies have the highest number of votes?

In [55]:

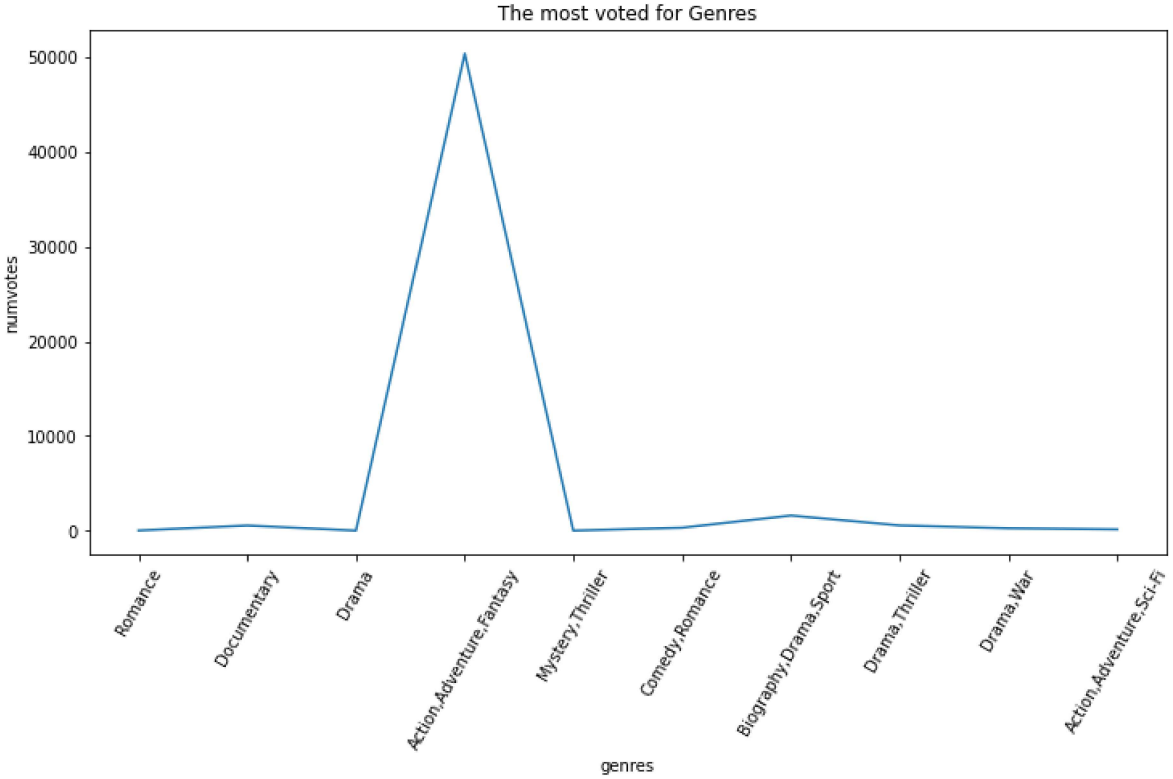
```
#checking for the needed columns first
basics_and_ratings
```

Out[55]:

	movie_id	averagerating	numvotes	primary_title	original_title	start_year	runtime_minutes
0	tt10356526	8.3	31	Laiye Je Yaarian	Laiye Je Yaarian	2019	117.0
1	tt10384606	8.9	559	Borderless	Borderless	2019	87.0
2	tt1042974	6.4	20	Just Inès	Just Inès	2010	90.0
3	tt1043726	4.2	50352	The Legend of Hercules	The Legend of Hercules	2014	99.0
4	tt1060240	6.5	21	Até Onde?	Até Onde?	2011	73.0
5	tt1069246	6.2	326	Habana Eva	Habana Eva	2010	106.0
6	tt1094666	7.0	1613	The Hammer	Hamill	2010	108.0
7	tt1130982	6.4	571	The Night Clerk	Avant l'aube	2011	104.0
8	tt1156528	7.2	265	Silent Sonata	Circus Fantasticus	2011	77.0
9	tt1161457	4.2	148	Vanquisher	The Vanquisher	2016	90.0

In [56]:

```
#Plotting a seaborn lineplot
plt.figure(figsize=(12,6))
sns.lineplot( x="genres", y="numvotes", data=basics_and_ratings,)
plt.title("The most voted for Genres") #labelling
plt.xticks(rotation = 60);
plt.show()
```



The most voted for genre is Action,Adventure,Fantasy followed by Biography,Drama,Sport.

Conclusion

1. A movie's average rating does not guarantee that it is a good movie, and the opposite is also true.
2. Film studios should provide many online and offline access methods for their content.
3. Fans of movies convey a different message about what they find appealing in movies.
4. According to the data provided, romantic films had longer runs than scary films. Films that are near to the hearts of the audience should receive more attention than those that frighten them, as the production budget also increases somewhat as a result.
5. It is necessary to conduct more research. To determine the amount of individuals who really see movies in theaters versus those who prefer to stream, surveys can be sent to owners of movie theaters, moviegoers, and internet respondents.
6. Depending on their genre and production costs, movies can make money both domestic and foreign.. Less people will watch it the worse the quality, and vice versa.

Recommendations

1. The dataframes displayed the various movie genres, the titles of the films, their budgets for production, and the respective domestic, international, and global box office receipts for the film studios. Despite having a global and worldwide audience, the languages employed in the films did not take into account other continents; for instance, there was no swahili-language film or actor. Therefore,accessibility of content in different markets When movies come out should be considered. Allow growth by giving everyone the chance to watch a new movie in every region of the world.
2. Major Markets to invest in: .Tv Licensing .Foreign distribution .Domestic Box Office .Physical Copy sales .Digital streaming & video on demand
3. Consider first going through the company planning process.
4. Work in all languages and with the more popular genres.