Microsoft Movie Analysis

Author: Gustavo Villagrana

Table of Contents

Overview

This project analyzes the types of films that are currently doing the best at the box office to help the head of Microsoft's new movie studio decide what type of films to create. Since this is Microsoft's first time creating films, it is critical that the best option is clearly identified in order to optimize this investment opportunity.

Business Problem

Microsoft wants to create a new movie studio to produce original content but needs help in identifying what type of films are performing best at the box office. By identifying the best performing films, Microsoft will be able to leverage its investment resources and maximize its profitability.

Data Understanding

The data used in this our analysis are primarily sourced from the IMBd web site, which is considered one of the best online movie databases available. Every movie has a unique ID, primary and original titles, runtime minutes, genres, and domestic/foreign gross earnings.

In [13]:

```
# Import standard packages
import pandas as pd
pd.options.display.float_format = '{:,.2f}'.format
import numpy as np

import matplotlib.pyplot as plt
import matplotlib as mpl
%matplotlib inline
import seaborn as sns
```

Title Basics data

In [14]:

```
# Title Basics data

title_basics_df = pd.read_csv('data/imdb.title.basics.csv.gz')
title_basics_df.rename(columns={'tconst': 'movie_id'}, inplace=True)
title_basics_df
```

Out[14]:

	movie_id	primary_title	original_title	start_year	runtime_minutes	genres
0	tt0063540	Sunghursh	Sunghursh	2013	175.00	Action,Crime,Drama
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.00	Biography,Drama
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.00	Drama
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	nan	Comedy,Drama
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.00	Comedy,Drama,Fantasy
...	...	...	...	...	...	...
146139	tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.00	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	116.00	NaN
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	nan	Documentary

146144 rows × 6 columns

In [15]:

```
title_basics_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146144 entries, 0 to 146143
Data columns (total 6 columns):
 #   Column           Non-Null Count   Dtype  
---  ------           --------------   -----  
 0   movie_id         146144 non-null  object 
 1   primary_title    146144 non-null  object 
 2   original_title   146123 non-null  object 
 3   start_year       146144 non-null  int64  
 4   runtime_minutes  114405 non-null  float64
 5   genres           140736 non-null  object 
dtypes: float64(1), int64(1), object(4)
memory usage: 6.7+ MB
```

Title Ratings data

In [16]:

```
# Understanding the data:
# imdb.title.ratings

title_ratings_df = pd.read_csv('data/imdb.title.ratings.csv.gz')
title_ratings_df.rename(columns={'tconst': 'movie_id'}, inplace=True)
title_ratings_df
```

Out[16]:

	movie_id	averagerating	numvotes
0	tt10356526	8.30	31
1	tt10384606	8.90	559
2	tt1042974	6.40	20
3	tt1043726	4.20	50352
4	tt1060240	6.50	21
...	...	...	...
73851	tt9805820	8.10	25
73852	tt9844256	7.50	24
73853	tt9851050	4.70	14
73854	tt9886934	7.00	5
73855	tt9894098	6.30	128

73856 rows × 3 columns

In [17]:

```
title_ratings_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 73856 entries, 0 to 73855
Data columns (total 3 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   movie_id       73856 non-null  object 
 1   averagerating  73856 non-null  float64
 2   numvotes       73856 non-null  int64  
dtypes: float64(1), int64(1), object(1)
memory usage: 1.7+ MB
```

Movie Gross data

In [22]:

```
# Movie Gross data

# domestic_gross is FLOAT type but foreign_gross is a STRING type



movie_gross_df = pd.read_csv('data/bom.movie_gross.csv.gz')
movie_gross_df
```

Out[22]:

	title	studio	domestic_gross	foreign_gross	year
0	Toy Story 3	BV	415,000,000.00	652000000	2010
1	Alice in Wonderland (2010)	BV	334,200,000.00	691300000	2010
2	Harry Potter and the Deathly Hallows Part 1	WB	296,000,000.00	664300000	2010
3	Inception	WB	292,600,000.00	535700000	2010

	title	studio	domestic_gross	foreign_gross	year
4	Shrek Forever After	P/DW	238,700,000.00	513900000	2010
...	...	...	...	...	...
3382	The Quake	Magn.	6,200.00	NaN	2018
3383	Edward II (2018 re-release)	FM	4,800.00	NaN	2018
3384	El Pacto	Sony	2,500.00	NaN	2018
3385	The Swan	Synergetic	2,400.00	NaN	2018
3386	An Actor Prepares	Grav.	1,700.00	NaN	2018

3387 rows × 5 columns

In [23]:

```
movie_gross_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
```

Movie Budget data

In [25]:

```
movie_budget_df = pd.read_csv("data/tn.movie_budgets.csv.gz")
movie_budget_df
```

Out[25]:

	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 [26]:

```
movie_budget_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5782 entries, 0 to 5781
Data columns (total 6 columns):
```

```
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   object
 4   domestic_gross     5782 non-null   object
 5   worldwide_gross    5782 non-null   object
dtypes: int64(1), object(5)
memory usage: 271.2+ KB
```

Preparing and Cleaning the Title and Ratings Data

In [27]:

```
# Joined Title Basics df and Title Ratings df ON movie_id

basics_with_ratings_df = title_basics_df.join(title_ratings_df.set_index('movie_id'), on
='movie_id')
basics_with_ratings_df
```

Out[27]:

	movie_id	primary_title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
0	tt0063540	Sunghursh	Sunghursh	2013	175.00	Action,Crime,Drama	7.00	77.00
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.00	Biography,Drama	7.20	43.00
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.00	Drama	6.90	4,517.00
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	nan	Comedy,Drama	6.10	13.00
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.00	Comedy,Drama,Fantasy	6.50	119.00
...	...	...	...	...	...	...	...	...
146139	tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.00	Drama	nan	nan
146140	tt9916622	Rodolpho Teóphilo - O Legado de um Pioneiro	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	nan	Documentary	nan	nan
146141	tt9916706	Dankyavar Danka	Dankyavar Danka	2013	nan	Comedy	nan	nan
146142	tt9916730	6 Gunn	6 Gunn	2017	116.00	NaN	nan	nan
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	nan	Documentary	nan	nan

146144 rows × 8 columns

In [28]:

```
basics_with_ratings_df.rename(columns={'primary_title': 'title'}, inplace=True)
basics_with_ratings_df
```

Out[28]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
0	tt0063540	Sunghursh	Sunghursh	2013	175.00	Action,Crime,Drama	7.00	77.00
1	tt0066787	One Day Before the Rainy Season	Ashad Ka Ek Din	2019	114.00	Biography,Drama	7.20	43.00
2	tt0069049	The Other Side of the Wind	The Other Side of the Wind	2018	122.00	Drama	6.90	4,517.00
3	tt0069204	Sabse Bada Sukh	Sabse Bada Sukh	2018	nan	Comedy,Drama	6.10	13.00
4	tt0100275	The Wandering Soap Opera	La Telenovela Errante	2017	80.00	Comedy,Drama,Fantasy	6.50	119.00
...	...	...	...	...	...	...	...	...
146139	tt9916538	Kuambil Lagi Hatiku	Kuambil Lagi Hatiku	2019	123.00	Drama	nan	nan
146140	tt9916622	Rodolpho Teóphilo - O Legado de um Pioneiro	Rodolpho Teóphilo - O Legado de um Pioneiro	2015	nan	Documentary	nan	nan
146141	tt9916706	Dankyavar Danka	Dankyavar Danka	2013	nan	Comedy	nan	nan
146142	tt9916730	6 Gunn	6 Gunn	2017	116.00	NaN	nan	nan
146143	tt9916754	Chico Albuquerque - Revelações	Chico Albuquerque - Revelações	2013	nan	Documentary	nan	nan

146144 rows × 8 columns

Preparing and Cleaning Movie Gross Data

In [29]:

```
movie_gross_sorted_df = movie_gross_df.sort_values(by=['domestic_gross'], ascending=Fals
e)
movie_gross_sorted_df
```

Out[29]:

	title	studio	domestic_gross	foreign_gross	year
1872	Star Wars: The Force Awakens	BV	936,700,000.00	1,131.6	2015
3080	Black Panther	BV	700,100,000.00	646900000	2018
3079	Avengers: Infinity War	BV	678,800,000.00	1,369.5	2018
1873	Jurassic World	Uni.	652,300,000.00	1,019.4	2015
727	Marvel's The Avengers	BV	623,400,000.00	895500000	2012
...	...	...	...	...	...
1975	Surprise - Journey To The West	AR	nan	49600000	2015
2392	Finding Mr. Right 2	CL	nan	114700000	2016
2468	Solace	LGP	nan	22400000	2016
2595	Viral	W/Dim.	nan	552000	2016
2825	Secret Superstar	NaN	nan	122000000	2017

3387 rows × 5 columns

In [30]:

```
# Remove commas from df['foreign_gross'] column

movie_gross_sorted_df['foreign_gross'].replace(',','', regex=True, inplace=True)
movie_gross_sorted_df
```

Out[30]:

	title	studio	domestic_gross	foreign_gross	year
1872	**Star Wars: The Force Awakens**	**BV**	**936,700,000.00**	**1131.6**	**2015**
3080	**Black Panther**	**BV**	**700,100,000.00**	**646900000**	**2018**
3079	**Avengers: Infinity War**	**BV**	**678,800,000.00**	**1369.5**	**2018**
1873	**Jurassic World**	**Uni.**	**652,300,000.00**	**1019.4**	**2015**
727	**Marvel's The Avengers**	**BV**	**623,400,000.00**	**895500000**	**2012**
...	**...**	**...**	**...**	**...**	**...**
1975	**Surprise - Journey To The West**	**AR**	**nan**	**49600000**	**2015**
2392	**Finding Mr. Right 2**	**CL**	**nan**	**114700000**	**2016**
2468	**Solace**	**LGP**	**nan**	**22400000**	**2016**
2595	**Viral**	**W/Dim.**	**nan**	**552000**	**2016**
2825	**Secret Superstar**	**NaN**	**nan**	**122000000**	**2017**

3387 rows × 5 columns

In [31]:

```
movie_gross_sorted_df['foreign_gross'] = movie_gross_sorted_df['foreign_gross'].astype(f
loat)
movie_gross_sorted_df
```

Out[31]:

	title	studio	domestic_gross	foreign_gross	year
1872	**Star Wars: The Force Awakens**	**BV**	**936,700,000.00**	**1,131.60**	**2015**
3080	**Black Panther**	**BV**	**700,100,000.00**	**646,900,000.00**	**2018**
3079	**Avengers: Infinity War**	**BV**	**678,800,000.00**	**1,369.50**	**2018**
1873	**Jurassic World**	**Uni.**	**652,300,000.00**	**1,019.40**	**2015**
727	**Marvel's The Avengers**	**BV**	**623,400,000.00**	**895,500,000.00**	**2012**
...	**...**	**...**	**...**	**...**	**...**
1975	**Surprise - Journey To The West**	**AR**	**nan**	**49,600,000.00**	**2015**
2392	**Finding Mr. Right 2**	**CL**	**nan**	**114,700,000.00**	**2016**
2468	**Solace**	**LGP**	**nan**	**22,400,000.00**	**2016**
2595	**Viral**	**W/Dim.**	**nan**	**552,000.00**	**2016**
2825	**Secret Superstar**	**NaN**	**nan**	**122,000,000.00**	**2017**

3387 rows × 5 columns

In [32]:

```
movie_gross_sorted_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3387 entries, 1872 to 2825
Data columns (total 5 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   title           3387 non-null   object
 1   studio          3382 non-null   object
```

```
 1   studio           3382 non-null   object
 2   domestic_gross   3359 non-null   float64
 3   foreign_gross    2037 non-null   float64
 4   year             3387 non-null   int64
dtypes: float64(2), int64(1), object(2)
memory usage: 158.8+ KB
```

In [33]:

```
movie_gross_sorted_df
```

Out[33]:

	title	studio	domestic_gross	foreign_gross	year
1872	Star Wars: The Force Awakens	BV	936,700,000.00	1,131.60	2015
3080	Black Panther	BV	700,100,000.00	646,900,000.00	2018
3079	Avengers: Infinity War	BV	678,800,000.00	1,369.50	2018
1873	Jurassic World	Uni.	652,300,000.00	1,019.40	2015
727	Marvel's The Avengers	BV	623,400,000.00	895,500,000.00	2012
...	...	...	...	...	...
1975	Surprise - Journey To The West	AR	nan	49,600,000.00	2015
2392	Finding Mr. Right 2	CL	nan	114,700,000.00	2016
2468	Solace	LGP	nan	22,400,000.00	2016
2595	Viral	W/Dim.	nan	552,000.00	2016
2825	Secret Superstar	NaN	nan	122,000,000.00	2017

3387 rows × 5 columns

Joining the Data

In [34]:

```
basics_ratings_gross_df = basics_with_ratings_df.join(movie_gross_sorted_df.set_index('title'), how='inner', on='title')
basics_ratings_gross_df
```

Out[34]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
38	tt0315642	Wazir	Wazir	2016	103.00	Action,Crime,Drama	7.10	15,378.00
48	tt0337692	On the Road	On the Road	2012	124.00	Adventure,Drama,Romance	6.10	37,886.00
39490	tt2404548	On the Road	On the Road	2011	90.00	Drama	nan	nan
68078	tt3872966	On the Road	On the Road	2013	87.00	Documentary	nan	nan
76007	tt4339118	On the Road	On the Road	2014	89.00	Drama	6.00	6.00
...	...	...	...	...	...	...	...	...
133797	tt8404272	How Long Will I Love U	Chao shi kong tong ju	2018	101.00	Romance	6.50	607.00
134045	tt8427036	Helicopter Eela	Helicopter Eela	2018	135.00	Drama	5.40	673.00
137854	tt8851262	Spring Fever	Spring Fever	2019	nan	Comedy,Horror	nan	nan

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
140171	tt9078374	Last Letter	Ni hao, Zhihua	2018	114.00	Drama,Romance	6.40	322.00
140826	tt9151704	Burn the Stage: The Movie	Burn the Stage: The Movie	2018	84.00	Documentary,Music	8.80	2,067.00

3366 rows × 12 columns

In [35]:

```
basics_ratings_gross_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3366 entries, 38 to 140826
Data columns (total 12 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   movie_id         3366 non-null   object 
 1   title            3366 non-null   object 
 2   original_title   3366 non-null   object 
 3   start_year       3366 non-null   int64  
 4   runtime_minutes  3198 non-null   float64
 5   genres           3326 non-null   object 
 6   averagerating    3027 non-null   float64
 7   numvotes         3027 non-null   float64
 8   studio           3363 non-null   object 
 9   domestic_gross   3342 non-null   float64
 10  foreign_gross    2043 non-null   float64
 11  year             3366 non-null   int64  
dtypes: float64(5), int64(2), object(5)
memory usage: 341.9+ KB
```

In [36]:

```
df_without_nan = basics_ratings_gross_df.dropna()
df_without_nan
```

Out[36]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvote
48	tt0337692	On the Road	On the Road	2012	124.00	Adventure,Drama,Romance	6.10	37,886.
76007	tt4339118	On the Road	On the Road	2014	89.00	Drama	6.00	6.
96791	tt5647250	On the Road	On the Road	2016	121.00	Drama	5.70	127.
54	tt0359950	The Secret Life of Walter Mitty	The Secret Life of Walter Mitty	2013	114.00	Adventure,Comedy,Drama	7.30	275,300.
58	tt0365907	A Walk Among the Tombstones	A Walk Among the Tombstones	2014	114.00	Action,Crime,Drama	6.50	105,116.
...	...	...	...	...	...	...	...	
126784	tt7752454	Detective Chinatown 2	Tang ren jie tan an 2	2018	121.00	Action,Comedy,Mystery	6.10	1,250.
127205	tt7784604	Hereditary	Hereditary	2018	127.00	Drama,Horror,Mystery	7.30	151,571.
130621	tt8097306	Nobody's Fool	Nobody's Fool	2018	110.00	Comedy,Drama,Romance	4.60	3,618.
133797	tt8404272	How Long Will I Love U	Chao shi kong tong ju	2018	101.00	Romance	6.50	607.
140826	tt9151704	Burn the Stage: The	Burn the Stage: The	2018	84.00	Documentary,Music	8.80	2,067.

	movie_id	Movie title	Movie original_title	start_year	runtime_minutes	genres	averagerating	numvot

1767 rows × 12 columns

In [37]:

```
# Add a new column 'total_gross' to df_without_nan df

df_without_nan['total_gross'] = df_without_nan['domestic_gross'] + df_without_nan['foreign_gross']
df_without_nan
```

```
<ipython-input-37-5c46eb2b29c7>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_without_nan['total_gross'] = df_without_nan['domestic_gross'] + df_without_nan['foreign_gross']
```

Out[37]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvot
48	tt0337692	On the Road	On the Road	2012	124.00	Adventure,Drama,Romance	6.10	37,886.
76007	tt4339118	On the Road	On the Road	2014	89.00	Drama	6.00	6.
96791	tt5647250	On the Road	On the Road	2016	121.00	Drama	5.70	127.
54	tt0359950	The Secret Life of Walter Mitty	The Secret Life of Walter Mitty	2013	114.00	Adventure,Comedy,Drama	7.30	275,300.
58	tt0365907	A Walk Among the Tombstones	A Walk Among the Tombstones	2014	114.00	Action,Crime,Drama	6.50	105,116.
...	...	...	...	...	...	...	...	
126784	tt7752454	Detective Chinatown 2	Tang ren jie tan an 2	2018	121.00	Action,Comedy,Mystery	6.10	1,250.
127205	tt7784604	Hereditary	Hereditary	2018	127.00	Drama,Horror,Mystery	7.30	151,571.
130621	tt8097306	Nobody's Fool	Nobody's Fool	2018	110.00	Comedy,Drama,Romance	4.60	3,618.
133797	tt8404272	How Long Will I Love U	Chao shi kong tong ju	2018	101.00	Romance	6.50	607.
140826	tt9151704	Burn the Stage: The Movie	Burn the Stage: The Movie	2018	84.00	Documentary,Music	8.80	2,067.

1767 rows × 13 columns

In [38]:

```
df_without_nan.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1767 entries, 48 to 140826
Data columns (total 13 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   movie_id         1767 non-null   object
 1   title            1767 non-null   object
 2   original_title   1767 non-null   object
 3   start_year       1767 non-null   int64
```

```
 4   runtime_minutes   1767 non-null   float64
 5   genres            1767 non-null   object
 6   averagerating     1767 non-null   float64
 7   numvotes          1767 non-null   float64
 8   studio            1767 non-null   object
 9   domestic_gross    1767 non-null   float64
 10  foreign_gross     1767 non-null   float64
 11  year              1767 non-null   int64
 12  total_gross       1767 non-null   float64
dtypes: float64(6), int64(2), object(5)
memory usage: 193.3+ KB
```

In [39]:

```
# Attempting to sort by total_gross by DESC

df_without_nan = df_without_nan.sort_values(by='total_gross', ascending=False)
df_without_nan.head(10)
```

Out[39]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvote
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.0
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,148.0
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.10	462,903.0
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.20	219,125.0
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.20	62,311.0
10824	tt1611845	Frozen	Wai nei chung ching	2010	92.00	Fantasy,Romance	5.40	75.0
35107	tt2294629	Frozen	Frozen	2013	102.00	Adventure,Animation,Comedy	7.50	516,998.0
62741	tt3606756	Incredibles 2	Incredibles 2	2018	118.00	Action,Adventure,Animation	7.70	203,510.0
6453	tt1300854	Iron Man 3	Iron Man Three	2013	130.00	Action,Adventure,Sci-Fi	7.20	692,794.0
35077	tt2293640	Minions	Minions	2015	91.00	Adventure,Animation,Comedy	6.40	193,917.0

Remove Duplicate Titles in Data

In [40]:

```
df_without_nan.original_title.duplicated().sum()    #used original_title to check duplicates
```

Out[40]:

```
156
```

In [41]:

```
# drop first duplicate using original_title column

df_no_duplicates = df_without_nan.drop_duplicates(subset=['original_title'])
df_no_duplicates
```

Out[41]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.3
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.3
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.1
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.2
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.2
...	...	...	...	...	...	...	..
112563	tt6599340	Bluebeard	Haebing	2017	117.00	Thriller	6.4
71753	tt4096620	Troublemakers: The Story of Land Art	Troublemakers: The Story of Land Art	2015	72.00	Biography,Documentary,History	6.5
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.2
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.3
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.3

1611 rows × 13 columns

Working with Genres data

In [42]:

```
df_no_duplicates
```

Out[42]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.3
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.3
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.1
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.2
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.2
...	...	...	...	...	...	...	..
112563	tt6599340	Bluebeard	Haebing	2017	117.00	Thriller	6.4
71753	tt4096620	Troublemakers: The Story of Land Art	Troublemakers: The Story of Land Art	2015	72.00	Biography,Documentary,History	6.5
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.2
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.3
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.3

1611 rows × 13 columns

In [43]:

```
df_no_duplicates['list_of_genres'] = df_no_duplicates['genres'].map(lambda x: x.split(',
'))
df_no_duplicates
```

```
<ipython-input-43-d089ba3f9c85>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_g
uide/indexing.html#returning-a-view-versus-a-copy
  df_no_duplicates['list_of_genres'] = df_no_duplicates['genres'].map(lambda x: x.split('
,'))
```

Out[43]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.3
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.3
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.1
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.2
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.2
...	...	...	...	...	...	...	..
112563	tt6599340	Bluebeard	Haebing	2017	117.00	Thriller	6.4
71753	tt4096620	Troublemakers: The Story of Land Art	Troublemakers: The Story of Land Art	2015	72.00	Biography,Documentary,History	6.5
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.2
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.3
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.3

1611 rows × 14 columns

In [44]:

```
genres_df = df_no_duplicates.explode('list_of_genres')
genres_df
```

Out[44]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,148.00
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,148.00
...	...	...	...	...	...	...	...	...
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.20	766.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.30	7.00

3995 rows × 14 columns

In [45]:

```
# Sort total_gross values


genres_df_sorted = genres_df.sort_values('total_gross', ascending=False)
genres_df_sorted
```

Out[45]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,594.00
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,148.00
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,148.00
...	...	...	...	...	...	...	...	...
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.20	766.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.30	2,336.00
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.30	7.00

3995 rows × 14 columns

Groupby List of Genres

In [46]:

```
genres_grouped = genres_df_sorted.groupby('list_of_genres').sum()
```

```
genres_grouped
```

Out[46]:

list_of_genres	start_year	runtime_minutes	averagerating	numvotes	domestic_gross	foreign_gross	year	
Action	865967	49,790.00	2,773.20	74,107,413.00	36,335,362,197.00	65,617,712,558.90	866018	101,953
Adventure	696861	38,625.00	2,260.80	64,608,953.00	39,918,354,795.00	74,818,794,086.90	696896	114,737
Animation	241659	11,490.00	798.00	11,687,346.00	12,457,110,498.00	23,787,117,999.00	241697	36,244
Biography	294083	17,064.00	1,028.80	14,280,676.00	5,087,068,799.00	6,571,554,400.00	294117	11,658
Comedy	1155678	59,833.00	3,599.50	46,965,129.00	29,141,552,295.00	43,345,483,997.00	1155787	72,487
Crime	461046	25,541.00	1,490.90	24,457,749.00	8,100,952,300.00	8,811,448,972.00	461084	16,912
Documentary	130850	5,634.00	464.60	591,222.00	1,318,455,099.00	1,452,949,500.00	130851	2,771
Drama	1697262	95,226.00	5,624.00	68,497,066.00	22,776,665,897.00	34,626,839,094.00	1697470	57,403
Family	159031	8,124.00	482.40	5,285,789.00	4,966,634,100.00	7,561,901,300.00	159029	12,528
Fantasy	243638	13,476.00	758.60	17,062,051.00	8,773,336,799.00	17,946,549,700.00	243663	26,719
History	151042	8,805.00	517.00	5,067,437.00	2,209,447,099.00	3,099,532,399.00	151071	5,308
Horror	271869	13,366.00	783.10	11,036,345.00	5,430,041,700.00	7,388,494,200.00	271876	12,818
Music	90617	4,805.00	297.70	3,128,644.00	1,420,621,200.00	1,947,456,000.00	90616	3,368
Musical	16106	1,002.00	48.60	430,386.00	378,249,000.00	632,285,300.00	16107	1,010
Mystery	245656	12,999.00	760.10	14,681,759.00	4,387,561,200.00	6,564,975,498.00	245672	10,952
News	2014	52.00	6.40	8.00	21,400,000.00	43,200,000.00	2011	64
Romance	485100	26,270.00	1,534.10	14,571,599.00	6,319,340,999.00	8,515,053,795.00	485148	14,834
Sci-Fi	207423	12,121.00	679.40	31,003,174.00	14,001,881,299.00	22,597,943,388.90	207435	36,599
Sport	54363	3,212.00	190.10	2,209,506.00	1,194,054,700.00	1,342,702,500.00	54361	2,536
Thriller	515425	27,931.00	1,612.20	31,272,803.00	11,758,820,100.00	19,802,292,870.00	515483	31,561
War	38250	2,215.00	131.10	599,389.00	201,082,800.00	365,316,000.00	38256	560
Western	20118	1,148.00	64.60	1,915,496.00	474,139,100.00	680,623,000.00	20115	1,154

In [47]:

```
# Need to reset index

genres_grouped.reset_index(inplace=True)
genres_grouped
```

Out[47]:

	list_of_genres	start_year	runtime_minutes	averagerating	numvotes	domestic_gross	foreign_gross	year	
0	Action	865967	49,790.00	2,773.20	74,107,413.00	36,335,362,197.00	65,617,712,558.90	866018	101
1	Adventure	696861	38,625.00	2,260.80	64,608,953.00	39,918,354,795.00	74,818,794,086.90	696896	114
2	Animation	241659	11,490.00	798.00	11,687,346.00	12,457,110,498.00	23,787,117,999.00	241697	36
3	Biography	294083	17,064.00	1,028.80	14,280,676.00	5,087,068,799.00	6,571,554,400.00	294117	11
4	Comedy	1155678	59,833.00	3,599.50	46,965,129.00	29,141,552,295.00	43,345,483,997.00	1155787	72
5	Crime	461046	25,541.00	1,490.90	24,457,749.00	8,100,952,300.00	8,811,448,972.00	461084	16
6	Documentary	130850	5,634.00	464.60	591,222.00	1,318,455,099.00	1,452,949,500.00	130851	2
7	Drama	1697262	95,226.00	5,624.00	68,497,066.00	22,776,665,897.00	34,626,839,094.00	1697470	57
8	Family	159031	8,124.00	482.40	5,285,789.00	4,966,634,100.00	7,561,901,300.00	159029	12
9	Fantasy	243638	13,476.00	758.60	17,062,051.00	8,773,336,799.00	17,946,549,700.00	243663	26

	list_of_genres	start_year	runtime_minutes	averagerating	numvotes	domestic_gross	foreign_gross	year	
10	History	151042	8,805.00	517.00	5,067,437.00	2,209,447,099.00	3,099,532,399.00	151071	5
11	Horror	271869	13,366.00	783.10	11,036,345.00	5,430,041,700.00	7,388,494,200.00	271876	12
12	Music	90617	4,805.00	297.70	3,128,644.00	1,420,621,200.00	1,947,456,000.00	90616	3
13	Musical	16106	1,002.00	48.60	430,386.00	378,249,000.00	632,285,300.00	16107	1
14	Mystery	245656	12,999.00	760.10	14,681,759.00	4,387,561,200.00	6,564,975,498.00	245672	10
15	News	2014	52.00	6.40	8.00	21,400,000.00	43,200,000.00	2011	
16	Romance	485100	26,270.00	1,534.10	14,571,599.00	6,319,340,999.00	8,515,053,795.00	485148	14
17	Sci-Fi	207423	12,121.00	679.40	31,003,174.00	14,001,881,299.00	22,597,943,388.90	207435	30
18	Sport	54363	3,212.00	190.10	2,209,506.00	1,194,054,700.00	1,342,702,500.00	54361	2
19	Thriller	515425	27,931.00	1,612.20	31,272,803.00	11,758,820,100.00	19,802,292,870.00	515483	31
20	War	38250	2,215.00	131.10	599,389.00	201,082,800.00	365,316,000.00	38256	
21	Western	20118	1,148.00	64.60	1,915,496.00	474,139,100.00	680,623,000.00	20115	1

Plot Genres vs Total Gross Earnings

In [48]:

```
top_10_movie_gross = genres_grouped.nlargest(10, 'total_gross')
top_10_movie_gross
```

Out[48]:

	list_of_genres	start_year	runtime_minutes	averagerating	numvotes	domestic_gross	foreign_gross	year	
1	Adventure	696861	38,625.00	2,260.80	64,608,953.00	39,918,354,795.00	74,818,794,086.90	696896	114
0	Action	865967	49,790.00	2,773.20	74,107,413.00	36,335,362,197.00	65,617,712,558.90	866018	101
4	Comedy	1155678	59,833.00	3,599.50	46,965,129.00	29,141,552,295.00	43,345,483,997.00	1155787	72
7	Drama	1697262	95,226.00	5,624.00	68,497,066.00	22,776,665,897.00	34,626,839,094.00	1697470	57
17	Sci-Fi	207423	12,121.00	679.40	31,003,174.00	14,001,881,299.00	22,597,943,388.90	207435	30
2	Animation	241659	11,490.00	798.00	11,687,346.00	12,457,110,498.00	23,787,117,999.00	241697	30
19	Thriller	515425	27,931.00	1,612.20	31,272,803.00	11,758,820,100.00	19,802,292,870.00	515483	31
9	Fantasy	243638	13,476.00	758.60	17,062,051.00	8,773,336,799.00	17,946,549,700.00	243663	20
5	Crime	461046	25,541.00	1,490.90	24,457,749.00	8,100,952,300.00	8,811,448,972.00	461084	10
16	Romance	485100	26,270.00	1,534.10	14,571,599.00	6,319,340,999.00	8,515,053,795.00	485148	14

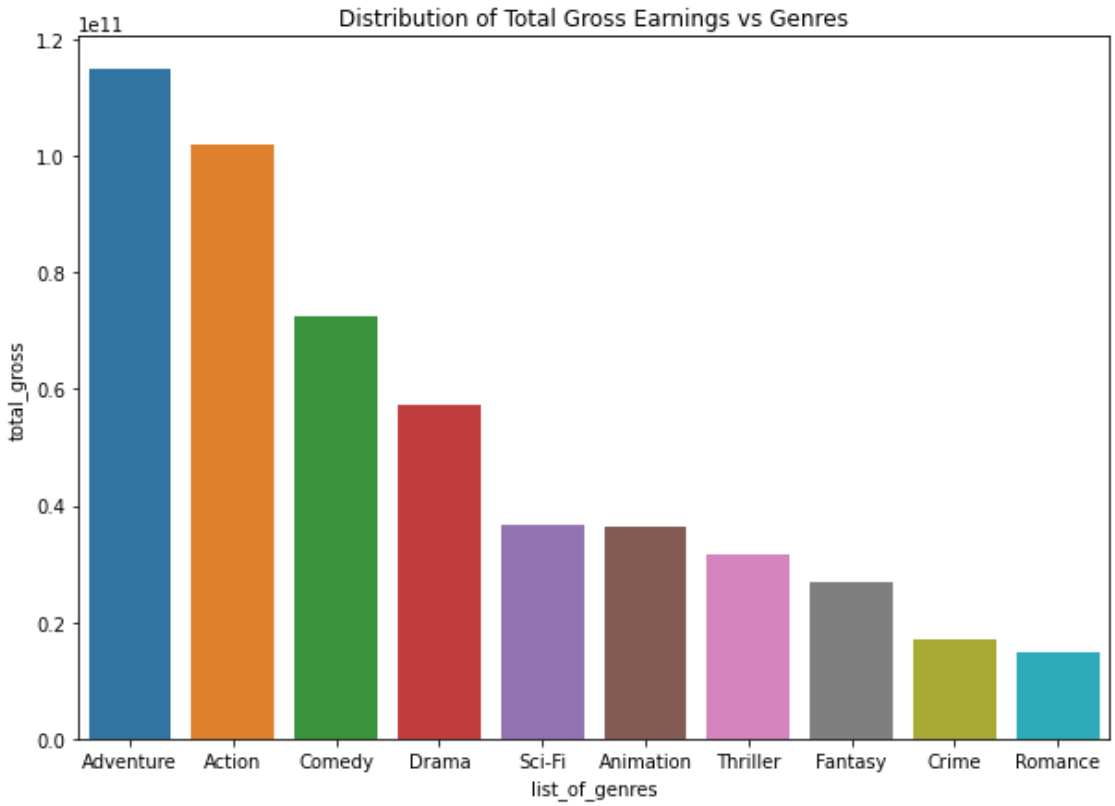
In [49]:

```
top_10_movie_gross.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 10 entries, 1 to 16
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   list_of_genres   10 non-null     object 
 1   start_year       10 non-null     int64  
 2   runtime_minutes  10 non-null     float64
 3   averagerating    10 non-null     float64
 4   numvotes         10 non-null     float64
 5   domestic_gross   10 non-null     float64
 6   foreign_gross    10 non-null     float64
 7   year             10 non-null     int64  
 8   total_gross      10 non-null     float64
dtypes: float64(6), int64(2), object(1)
memory usage: 800.0+ bytes
```

In [52]:

```
# Need to create a plot using genres and total_gross

plt.figure(figsize=(10, 7))

sns.barplot(
            x='list_of_genres',
            y='total_gross',
            data=top_10_movie_gross);

plt.title('Distribution of Total Gross Earnings vs Genres');
```



Working with Runtime Minutes data

In [53]:

```
run_time_df = df_no_duplicates.copy()
run_time_df
```

Out[53]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.3
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.3
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.1
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.2
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.2
...	...	...	...	...	...	...	...
112563	tt6599340	Bluebeard	Haebing	2017	117.00	Thriller	6.4

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
71753	tt4096620	The Story of Land Art	The Story of Land Art	2015	72.00	Biography,Documentary,History	6.5
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.2
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.3
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.3

1611 rows × 14 columns

In [54]:

```
max_runtime_df = run_time_df.sort_values(by=['runtime_minutes'], ascending=False)
max_runtime_df
```

Out[54]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvotes
6067	tt1236371	Mysteries of Lisbon	Mistérios de Lisboa	2010	272.00	Drama,Mystery,Romance	7.50	2,928.00
56699	tt3313066	Coriolanus	National Theatre Live: Coriolanus	2014	192.00	Drama,History,War	8.70	1,347.00
73161	tt4169250	M.S. Dhoni: The Untold Story	M.S. Dhoni: The Untold Story	2016	184.00	Biography,Drama,Sport	7.70	28,343.00
100712	tt5886728	Another Year	You yi nian	2016	181.00	Documentary	7.20	40.00
34528	tt2278871	Blue Is the Warmest Color	La vie d'Adèle	2013	180.00	Drama,Romance	7.80	124,409.00
...	...	...	...	...	...	...	...	...
38579	tt2385006	Jerusalem	Jerusalem	2013	44.00	Documentary	7.30	1,056.00
8837	tt1529567	Sea Rex 3D: Journey to a Prehistoric World	Sea Rex 3D: Journey to a Prehistoric World	2010	41.00	Documentary	6.90	364.00
45582	tt2713406	Meerkats	Meerkats	2011	40.00	Documentary	7.40	7.00
54321	tt3195742	Journey to the South Pacific	Journey to the South Pacific	2013	40.00	Documentary	6.50	145.00
49175	tt2926868	The Call	Lokroep	2013	25.00	Documentary	7.90	12.00

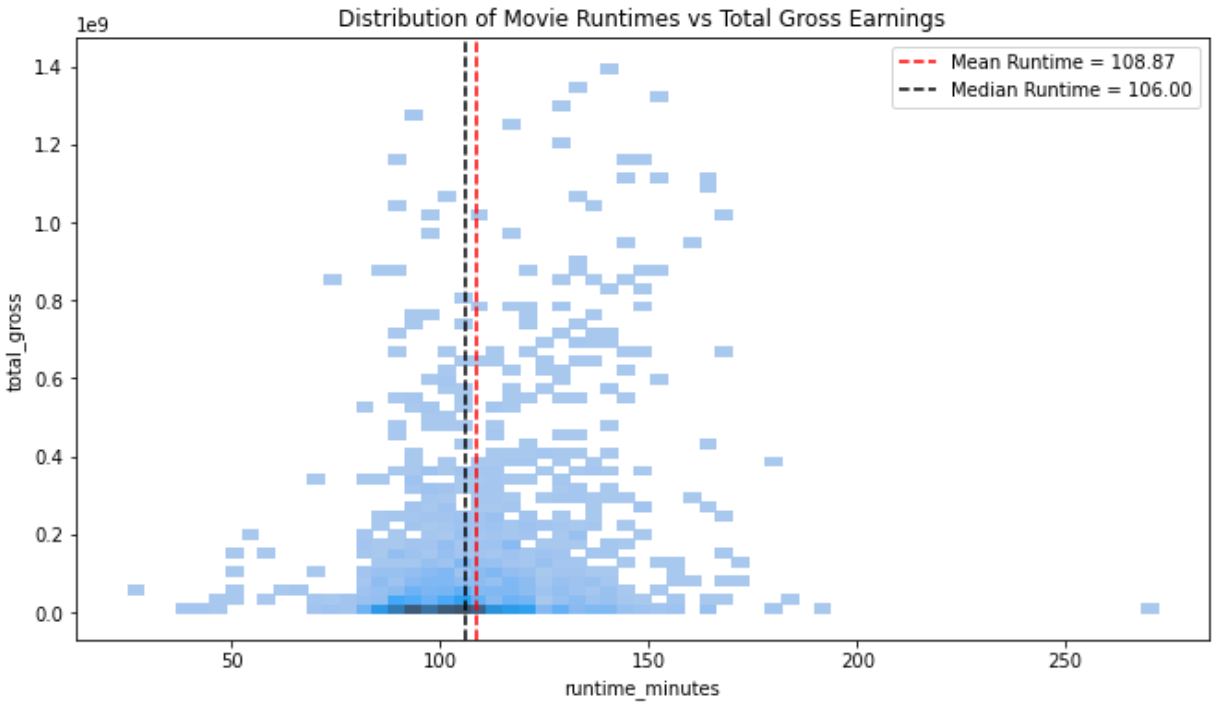
1611 rows × 14 columns

Plot Runtimes vs Total Gross Earnings

In [55]:

```
fig, ax = plt.subplots(figsize=(11, 6))

sns.histplot(data=max_runtime_df, x='runtime_minutes',
             y='total_gross', bins='auto')
```

```
ax.set(title='Distribution of Movie Runtimes vs Total Gross Earnings');

mean_runtime = round(max_runtime_df['runtime_minutes'].mean(),2)
ax.axvline(mean_runtime, color='red', ls='--',
           label=f"Mean Runtime = {mean_runtime:,.2f}");

median_runtime = round(max_runtime_df['runtime_minutes'].median(),2)
ax.axvline(median_runtime, color='black', ls='--',
           label=f"Median Runtime = {median_runtime:,.2f}");

ax.legend();
```



Working with Movie Budgets data

In [56]:

```
movie_prod_no_dups = df_no_duplicates.copy()
movie_prod_no_dups
```

Out[56]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
39010	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.3
19050	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.3
42223	tt2527336	Star Wars: The Last Jedi	Star Wars: Episode VIII - The Last Jedi	2017	152.00	Action,Adventure,Fantasy	7.1
84414	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.2
6647	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.2
...	...	...	...	...	...	...	...
112563	tt6599340	Bluebeard	Haebing	2017	117.00	Thriller	6.4
71753	tt4096620	Troublemakers: The Story of Land Art	Troublemakers: The Story of Land Art	2015	72.00	Biography,Documentary,History	6.5

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating
24172	tt1978447	Policeman	Ha-shoter	2011	105.00	Drama	6.2
7416	tt1417067	Cirkus Columbia	Cirkus Columbia	2010	113.00	Comedy,Drama,Romance	7.3
133711	tt8396182	Aurora	Aurora	2018	98.00	Drama	6.3

1611 rows × 14 columns

In [57]:

```
# Calling original movie_budget data

movie_budget_df
```

Out[57]:

	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 [58]:

```
# Rename 'movie' column to 'original_title' to match column name on both DFs.

movie_budget_df.rename(columns={'movie': 'original_title'}, inplace=True)
movie_budget_df
```

Out[58]:

	id	release_date	original_title	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 [59]:

```
# Drop columns from movie_budget_df that are NOT needed for the JOIN with
# no_duplicates_df.

movie_budget_dropped = movie_budget_df.drop(columns=['release_date', 'domestic_gross', '
worldwide_gross'])
movie_budget_dropped
```

Out[59]:

	id	original_title	production_budget
0	1	Avatar	$425,000,000
1	2	Pirates of the Caribbean: On Stranger Tides	$410,600,000
2	3	Dark Phoenix	$350,000,000
3	4	Avengers: Age of Ultron	$330,600,000
4	5	Star Wars Ep. VIII: The Last Jedi	$317,000,000
...	...	...	...
5777	78	Red 11	$7,000
5778	79	Following	$6,000
5779	80	Return to the Land of Wonders	$5,000
5780	81	A Plague So Pleasant	$1,400
5781	82	My Date With Drew	$1,100

5782 rows × 3 columns

In [60]:

```
# Need to remove '$' production_budget column
movie_budget_dropped['production_budget'] = movie_budget_dropped['production_budget'].str
.replace('$','')
movie_budget_dropped['production_budget']
```

Out[60]:

```
0       425,000,000
1       410,600,000
2       350,000,000
3       330,600,000
4       317,000,000
           ...
5777          7,000
5778          6,000
5779          5,000
5780          1,400
5781          1,100
Name: production_budget, Length: 5782, dtype: object
```

In [64]:

```
# Need to remove ',' from production_budget column
movie_budget_dropped['production_budget'] = movie_budget_dropped['production_budget'].str
.replace(',','')
movie_budget_dropped['production_budget']
```

Out[64]:

```
0       425000000
1       410600000
2       350000000
3       330600000
4       317000000
```

```
          ...
5777          7000
5778          6000
5779          5000
5780          1400
5781          1100
Name: production_budget, Length: 5782, dtype: object
```

In [65]:

```
# Conver string type to float

movie_budget_dropped['production_budget'] = movie_budget_dropped['production_budget'].ast
ype(float)
movie_budget_dropped
```

Out[65]:

	id	original_title	production_budget
0	1	Avatar	425,000,000.00
1	2	Pirates of the Caribbean: On Stranger Tides	410,600,000.00
2	3	Dark Phoenix	350,000,000.00
3	4	Avengers: Age of Ultron	330,600,000.00
4	5	Star Wars Ep. VIII: The Last Jedi	317,000,000.00
...	...	...	...
5777	78	Red 11	7,000.00
5778	79	Following	6,000.00
5779	80	Return to the Land of Wonders	5,000.00
5780	81	A Plague So Pleasant	1,400.00
5781	82	My Date With Drew	1,100.00

5782 rows × 3 columns

In [66]:

```
# JOIN movie_prod_no_dups and revised movie_budget_dropped and assign to movie_production
_df.

movie_production_df = pd.merge(movie_prod_no_dups, movie_budget_dropped, how='inner', on
='original_title')
movie_production_df
```

Out[66]:

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvo
0	tt2395427	Avengers: Age of Ultron	Avengers: Age of Ultron	2015	141.00	Action,Adventure,Sci-Fi	7.30	665,59
1	tt1825683	Black Panther	Black Panther	2018	134.00	Action,Adventure,Sci-Fi	7.30	516,14
2	tt4881806	Jurassic World: Fallen Kingdom	Jurassic World: Fallen Kingdom	2018	128.00	Action,Adventure,Sci-Fi	6.20	219,12
3	tt1323045	Frozen	Frozen	2010	93.00	Adventure,Drama,Sport	6.20	62,31
4	tt3606756	Incredibles 2	Incredibles 2	2018	118.00	Action,Adventure,Animation	7.70	203,51
...	...	...	...	...	...	...	...	

	movie_id	title	original_title	start_year	runtime_minutes	genres	averagerating	numvo
1003	tt1183923	Welcome to the Rileys	Welcome to the Rileys	2010	110.00	Drama	7.00	22,21
1004	tt2387589	The Girl on the Train	The Girl on the Train	2013	80.00	Thriller	4.40	81
1005	tt6333056	City of Ghosts	City of Ghosts	2017	92.00	Documentary,War	7.40	2,92
1006	tt1788391	Kill List	Kill List	2011	95.00	Crime,Drama,Horror	6.40	32,80
1007	tt1152822	Freakonomics	Freakonomics	2010	93.00	Documentary	6.40	6,46

1008 rows × 16 columns

Plot Total Gross vs Budget and Genres

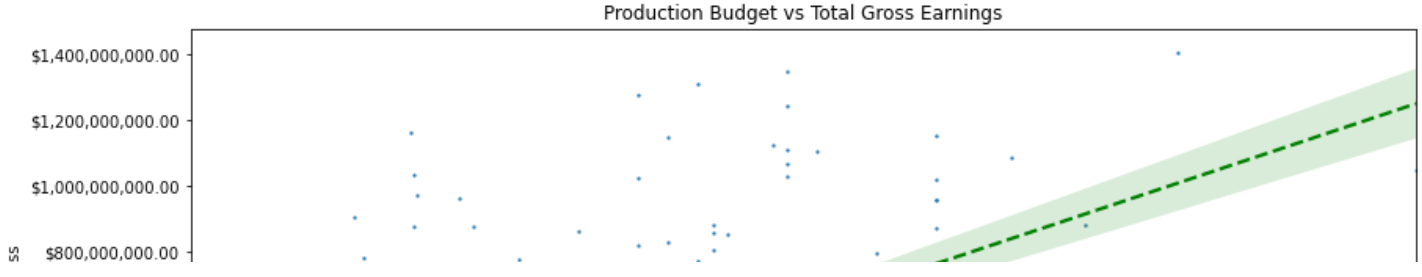
In [67]:

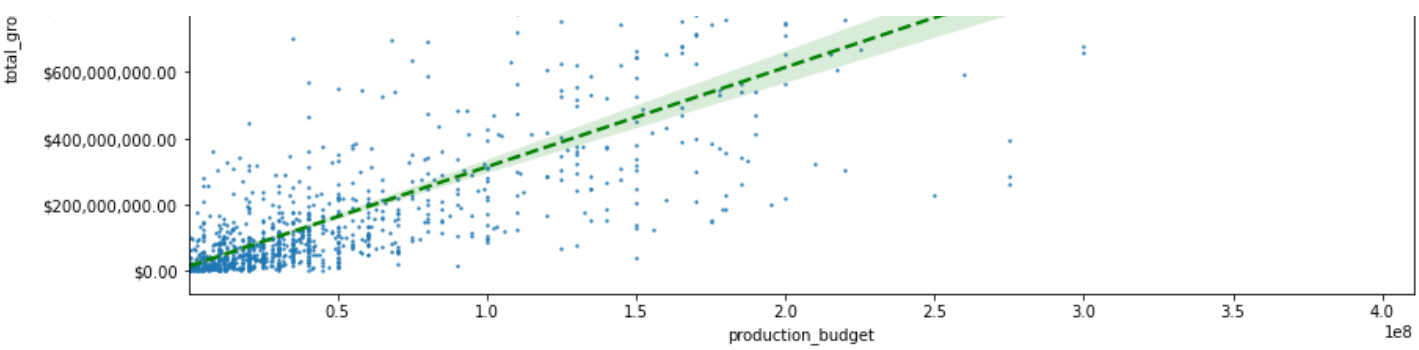
```
movie_production_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1008 entries, 0 to 1007
Data columns (total 16 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   movie_id           1008 non-null   object 
 1   title              1008 non-null   object 
 2   original_title     1008 non-null   object 
 3   start_year         1008 non-null   int64  
 4   runtime_minutes    1008 non-null   float64
 5   genres             1008 non-null   object 
 6   averagerating      1008 non-null   float64
 7   numvotes           1008 non-null   float64
 8   studio             1008 non-null   object 
 9   domestic_gross     1008 non-null   float64
 10  foreign_gross      1008 non-null   float64
 11  year               1008 non-null   int64  
 12  total_gross        1008 non-null   float64
 13  list_of_genres     1008 non-null   object 
 14  id                 1008 non-null   int64  
 15  production_budget  1008 non-null   float64
dtypes: float64(7), int64(3), object(6)
memory usage: 133.9+ KB
```

In [69]:

```
# Relationship between Total Gross and Production Budget and Genres

fig, ax = plt.subplots(figsize=(14,6))

price_fmt = mpl.ticker.StrMethodFormatter('${x:,.2f}')

# Plot 0
sns.regplot(data=movie_production_df, x='production_budget', y='total_gross',
            ax=ax, scatter_kws={'s':2}, line_kws={'color':'green', 'ls':'--'})

ax.set_title('Production Budget vs Total Gross Earnings')
ax.yaxis.set_major_formatter(price_fmt)
```





Conclusions

- **Microsoft should create Adventure films as that genre has the highest total gross earnings.**
- **Limit runtime range between 125 minutes - 150 minutes.**
- **Set production budget range between 150mm - 250mm to maximize total gross earnings.**

Next Steps

- **Calculate actual ROI results to determine a more accurate budget range.**