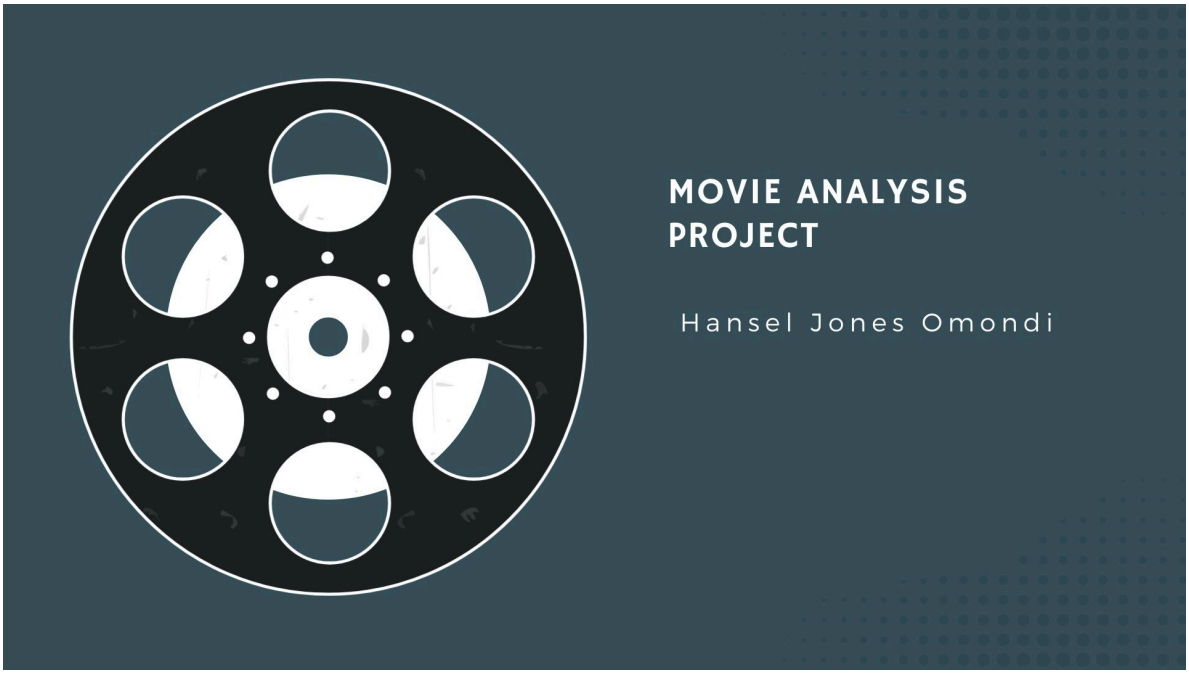
Final Project Submission

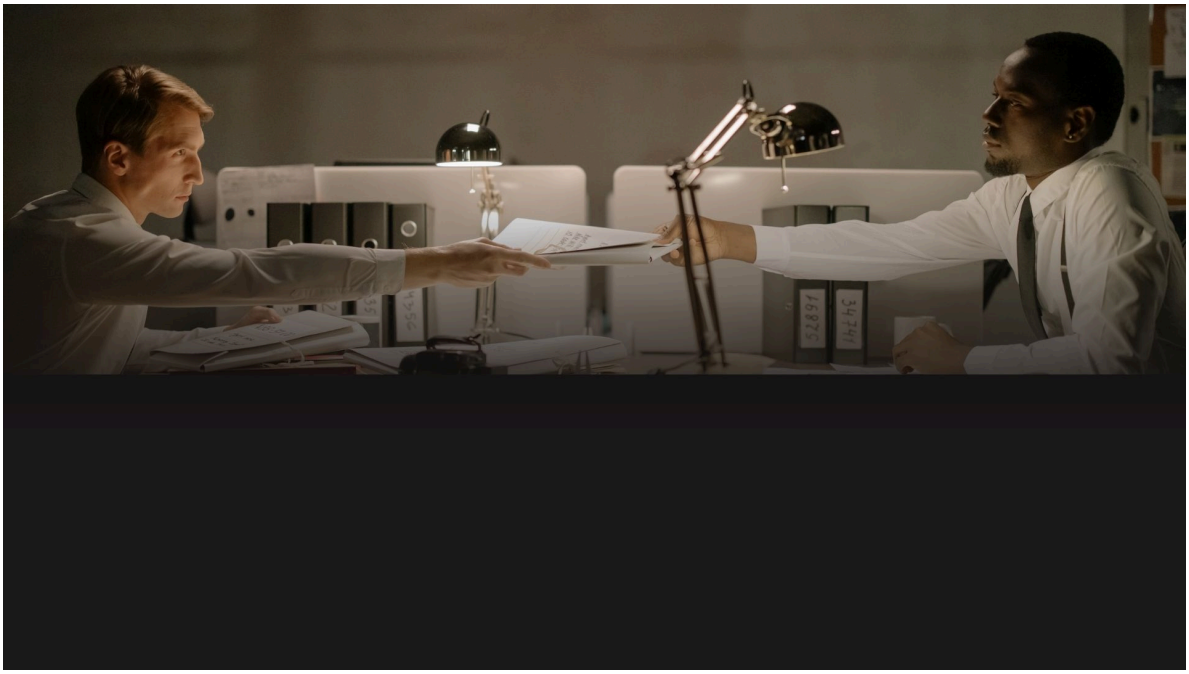


OVERVIEW

The company now desires to venture into creating a new movie studio and this project looks forward to explore the types of films that are currently doing the best at the box office after which thr project will translate those findings into actionable insights that the head of the company's new movie studio can use to help decide what type of films to create.

The goal of this project is to identify key factors that contribute to box office success, such as genre, budget, and review scores, and leverage these insights to make informed decisions about the types of films to produce.

DATA UNDERSTANDING



The movie datasets used contain the following;

- Rotten Tomatoes Reviews (rt.reviews.tsv.gz): Contains reviews and ratings for various movies
- The Numbers Movie Budgets (tn.movie_budgets.csv.gz): Includes production budgets and box office grosses
- Box Office Mojo Gross (bom.movie_gross.csv.gz): Details on box office grosses
- Rotten Tomatoes Movie Info (rt.movie_info.tsv.gz): Information about movie genres and release dates

DATA ANALYSIS AND DATA CLEANING

In [6]:
```
import pandas as pd
```

In [8]:

```
# Load datasets with specified encoding and delimiter
rt_reviews = pd.read_csv(r"C:\Users\Administrator\Documents\course_materia
tn_movie_budgets = pd.read_csv(r"C:\Users\Administrator\Documents\course_m
bom_movie_gross = pd.read_csv(r"C:\Users\Administrator\Documents\course_ma
rt_movie_info = pd.read_csv(r"C:\Users\Administrator\Documents\course_mate

# A Display of the first few rows of each dataset
print(rt_reviews.head())
print(tn_movie_budgets.head())
print(bom_movie_gross.head())
print(rt_movie_info.head())
```

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

          critic  top_critic         publisher               date
0      PJ Nabarro           0   Patrick Nabarro  November 10, 2018
1  Annalee Newitz           0           io9.com       May 23, 2018
2    Sean Axmaker           0  Stream on Demand    January 4, 2018
3   Daniel Kasman           0              MUBI  November 16, 2017
4             NaN           0      Cinema Scope   October 12, 2017
   id  release_date                                        movie  \
0   1  Dec 18, 2009                                       Avatar
1   2  May 20, 2011  Pirates of the Caribbean: On Stranger Tides
2   3   Jun 7, 2019                                  Dark Phoenix
3   4   May 1, 2015                      Avengers: Age of Ultron
4   5  Dec 15, 2017            Star Wars Ep. VIII: The Last Jedi

  production_budget domestic_gross worldwide_gross
0      $425,000,000   $760,507,625  $2,776,345,279
1      $410,600,000   $241,063,875  $1,045,663,875
2      $350,000,000    $42,762,350    $149,762,350
3      $330,600,000   $459,005,868  $1,403,013,963
4      $317,000,000   $620,181,382  $1,316,721,747
                                     title studio  domestic_gross  \
0                                Toy Story 3     BV     415000000.0
1                  Alice in Wonderland (2010)     BV     334200000.0
2  Harry Potter and the Deathly Hallows Part 1     WB     296000000.0
3                                  Inception     WB     292600000.0
4                          Shrek Forever After   P/DW     238700000.0

  foreign_gross  year
0     652000000  2010
1     691300000  2010
2     664300000  2010
3     535700000  2010
4     513900000  2010
   id                                          synopsis rating  \
0   1  This gritty, fast-paced, and innovative police...      R
1   3  New York City, not-too-distant-future: Eric Pa...      R
2   5  Illeana Douglas delivers a superb performance ...      R
3   6  Michael Douglas runs afoul of a treacherous su...      R
4   7                                                NaN     NR

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

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

```
1     David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013        $
2                   Allison Anders  Sep 13, 1996  Apr 18, 2000      NaN
3  Paul Attanasio|Michael Crichton   Dec 9, 1994  Aug 27, 1997      NaN
4                     Giles Cooper           NaN           NaN      NaN

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

In [9]:

```
# Converting financial columns to numerical values
tn_movie_budgets['production_budget'] = tn_movie_budgets['production_budge
tn_movie_budgets['domestic_gross'] = tn_movie_budgets['domestic_gross'].re
tn_movie_budgets['worldwide_gross'] = tn_movie_budgets['worldwide_gross'].
```

In [10]:

```
bom_movie_gross['domestic_gross'] = bom_movie_gross['domestic_gross'].repl
bom_movie_gross['foreign_gross'] = bom_movie_gross['foreign_gross'].replac
```

In [11]:

```
# Convertingrelease_date in tn_movie_budgets to datetime
tn_movie_budgets['release_date'] = pd.to_datetime(tn_movie_budgets['releas
```

In [13]:

```
# Check for missing values in each dataset
missing_values_rt_reviews = rt_reviews.isnull().sum()
missing_values_tn_movie_budgets = tn_movie_budgets.isnull().sum()
missing_values_bom_movie_gross = bom_movie_gross.isnull().sum()
missing_values_rt_movie_info = rt_movie_info.isnull().sum()

print("Missing values in Rotten Tomatoes Reviews:")
print(missing_values_rt_reviews)
print("\nMissing values in The Numbers Movie Budgets:")
print(missing_values_tn_movie_budgets)
print("\nMissing values in Box Office Mojo Movie Gross:")
print(missing_values_bom_movie_gross)
print("\nMissing values in Rotten Tomatoes Movie Info:")
print(missing_values_rt_movie_info)
```

```
Missing values in Rotten Tomatoes Reviews:
id                0
review         5563
rating        13517
fresh             0
critic         2722
top_critic        0
publisher       309
date              0
dtype: int64

Missing values in The Numbers Movie Budgets:
id                   0
release_date         0
movie                0
production_budget    0
domestic_gross       0
worldwide_gross      0
dtype: int64

Missing values in Box Office Mojo Movie Gross:
title                0
studio               5
domestic_gross      28
foreign_gross     1350
year                 0
dtype: int64

Missing values in Rotten Tomatoes Movie Info:
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 [14]:

```
# Dropping rows with missing values
rt_reviews_cleaned = rt_reviews.dropna()
tn_movie_budgets_cleaned = tn_movie_budgets.dropna()
bom_movie_gross_cleaned = bom_movie_gross.dropna()
rt_movie_info_cleaned = rt_movie_info.dropna()
```

In [16]:

```
# Displaying the cleaned datasets
print("Cleaned Rotten Tomatoes Reviews:")
print(rt_reviews_cleaned.head())
print("\nCleaned The Numbers Movie Budgets:")
print(tn_movie_budgets_cleaned.head())
print("\nCleaned Box Office Mojo Movie Gross:")
print(bom_movie_gross_cleaned.head())
print("\nCleaned Rotten Tomatoes Movie Info:")
print(rt_movie_info_cleaned.head())
```

```
Cleaned Rotten Tomatoes Reviews:
    id                                            review rating   fresh
\
0    3  A distinctly gallows take on contemporary fina...    3/5   fresh
6    3  Quickly grows repetitive and tiresome, meander...      C  rotten
7    3  Cronenberg is not a director to be daunted by ...    2/5  rotten
11   3  While not one of Cronenberg's stronger films, ...     B-   fresh
12   3  Robert Pattinson works mighty hard to make Cos...    2/4  rotten

            critic  top_critic           publisher               date
0       PJ Nabarro           0     Patrick Nabarro  November 10, 2018
6   Eric D. Snider           0     EricDSnider.com      July 17, 2013
7     Matt Kelemen           0  Las Vegas CityLife     April 21, 2013
11    Emanuel Levy           0     EmanuelLevy.Com   February 3, 2013
12  Christian Toto           0       Big Hollywood   January 15, 2013

Cleaned The Numbers Movie Budgets:
   id release_date                                       movie  \
0   1   2009-12-18                                      Avatar
1   2   2011-05-20  Pirates of the Caribbean: On Stranger Tides
2   3   2019-06-07                                 Dark Phoenix
3   4   2015-05-01                      Avengers: Age of Ultron
4   5   2017-12-15            Star Wars Ep. VIII: The Last Jedi

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

Cleaned Box Office Mojo Movie Gross:
                                       title studio  domestic_gross  \
0                                Toy Story 3     BV     415000000.0
1                  Alice in Wonderland (2010)     BV     334200000.0
2  Harry Potter and the Deathly Hallows Part 1     WB     296000000.0
3                                  Inception     WB     292600000.0
4                          Shrek Forever After   P/DW     238700000.0

   foreign_gross  year
0    652000000.0  2010
1    691300000.0  2010
2    664300000.0  2010
3    535700000.0  2010
4    513900000.0  2010

Cleaned Rotten Tomatoes Movie Info:
    id                                           synopsis rating  \
1    3  New York City, not-too-distant-future: Eric Pa...      R
6   10  Some cast and crew from NBC's highly acclaimed...  PG-13
7   13  Stewart Kane, an Irishman living in the Austra...      R
15  22  Two-time Academy Award Winner Kevin Spacey giv...      R
18  25  From ancient Japan's most enduring tale, the e...  PG-13

                                                genre             directo
r  \
1                   Drama|Science Fiction and Fantasy     David Cronenber
```

```
g
6                                                Comedy          Jake Kasda
n
7                                                 Drama         Ray Lawrenc
e
15                   Comedy|Drama|Mystery and Suspense  George Hickenloope
r
18  Action and Adventure|Drama|Science Fiction and...     Carl Erik Rinsc
h

                              writer  theater_date      dvd_date currency
\
1       David Cronenberg|Don DeLillo  Aug 17, 2012   Jan 1, 2013        $
6                         Mike White  Jan 11, 2002  Jun 18, 2002        $
7   Raymond Carver|Beatrix Christian  Apr 27, 2006   Oct 2, 2007        $
15                      Norman Snider  Dec 17, 2010   Apr 5, 2011        $
18        Chris Morgan|Hossein Amini  Dec 25, 2013   Apr 1, 2014        $

    box_office      runtime                  studio
1      600,000  108 minutes       Entertainment One
6   41,032,915   82 minutes      Paramount Pictures
7      224,114  123 minutes  Sony Pictures Classics
15   1,039,869  108 minutes            ATO Pictures
18  20,518,224  127 minutes      Universal Pictures
```

In [25]:

```
# Savin cleaned datasets to new files (optional)
rt_reviews_cleaned.to_csv('rt_reviews_cleaned.tsv', sep='\t', index=False)
tn_movie_budgets_cleaned.to_csv('tn_movie_budgets_cleaned.csv', index=Fals
bom_movie_gross_cleaned.to_csv('bom_movie_gross_cleaned.csv', index=False)
rt_movie_info_cleaned.to_csv('rt_movie_info_cleaned.tsv', sep='\t', index=
```

In [26]:

```
print(rt_movie_info_cleaned.columns)
print(bom_movie_gross_cleaned.columns)
```

```
Index(['id', 'synopsis', 'rating', 'genre', 'director', 'writer',
       'theater_date', 'dvd_date', 'currency', 'box_office', 'runtime',
       'studio'],
      dtype='object')
Index(['title', 'studio', 'domestic_gross', 'foreign_gross', 'year'], dty
pe='object')
```

In [20]:

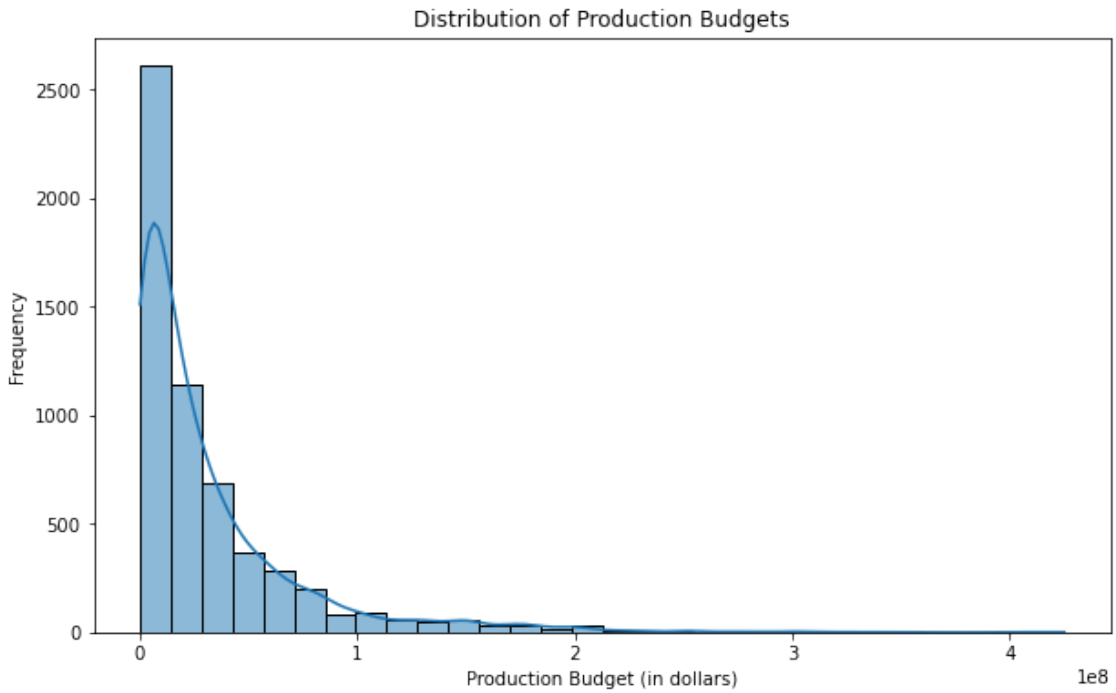
```
#Loading the required tools
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [21]:

```
# Loading cleaned datasets
rt_reviews_cleaned = pd.read_csv('rt_reviews_cleaned.tsv', delimiter='\t')
tn_movie_budgets_cleaned = pd.read_csv('tn_movie_budgets_cleaned.csv')
bom_movie_gross_cleaned = pd.read_csv('bom_movie_gross_cleaned.csv')
rt_movie_info_cleaned = pd.read_csv('rt_movie_info_cleaned.tsv', delimiter
```

Distribution of Production Budgets:

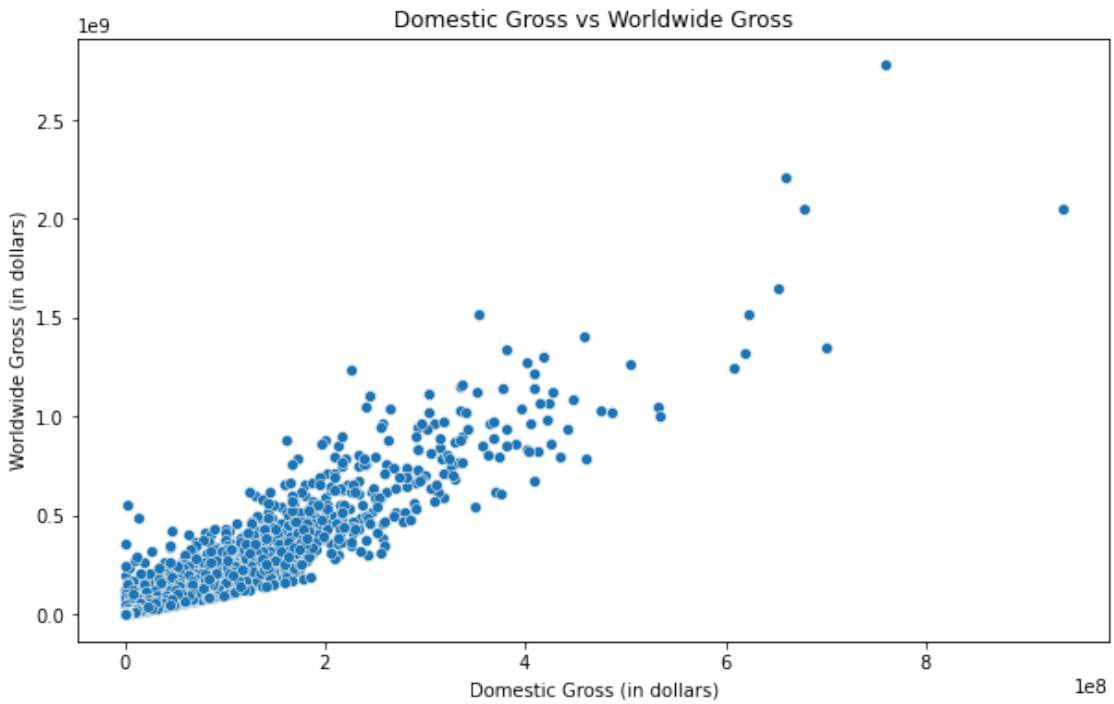
In [22]:
```
# VIZ 1: Distribution of Production Budgets
plt.figure(figsize=(10, 6))
sns.histplot(tn_movie_budgets_cleaned['production_budget'], bins=30, kde=T
plt.title('Distribution of Production Budgets')
plt.xlabel('Production Budget (in dollars)')
plt.ylabel('Frequency')
plt.show()
```



The distribution of production budgets shows that most movies have moderate budgets, with a few outliers having very high budgets. Recommendation: Since this is a new studio, starting with moderate-budget films might be wise. It balances the risk while still allowing for quality production.

Domestic vs Worldwide Gross:

In [23]:
```
# VIZ 2: Domestic vs Worldwide Gross
plt.figure(figsize=(10, 6))
sns.scatterplot(data=tn_movie_budgets_cleaned, x='domestic_gross', y='worl
plt.title('Domestic Gross vs Worldwide Gross')
plt.xlabel('Domestic Gross (in dollars)')
plt.ylabel('Worldwide Gross (in dollars)')
plt.show()
```



There is a positive correlation between domestic gross and worldwide gross. Successful films locally as per the visualization above tend to perform well internationally. Recommendation: Focus on producing films with universal appeal that can attract both domestic and international audiences to maximize revenue. A suggestion would be to consider the movies with languages that are globally recognized like English and Spanish.

In [36]:
```
print(rt_movie_info_cleaned.columns)
print(rt_movie_info_cleaned.dtypes)
print(bom_movie_gross_cleaned.columns)
print(bom_movie_gross_cleaned.dtypes)
```

```
Index(['id', 'synopsis', 'rating', 'genre', 'director', 'writer',
       'theater_date', 'dvd_date', 'currency', 'box_office', 'runtime',
       'studio'],
      dtype='object')
id               int64
synopsis        object
rating          object
genre           object
director        object
writer          object
theater_date    object
dvd_date        object
currency        object
box_office      object
runtime         object
studio          object
dtype: object
Index(['title', 'studio', 'domestic_gross', 'foreign_gross', 'year'], dty
pe='object')
title              object
studio             object
domestic_gross    float64
foreign_gross     float64
year                int64
dtype: object
```

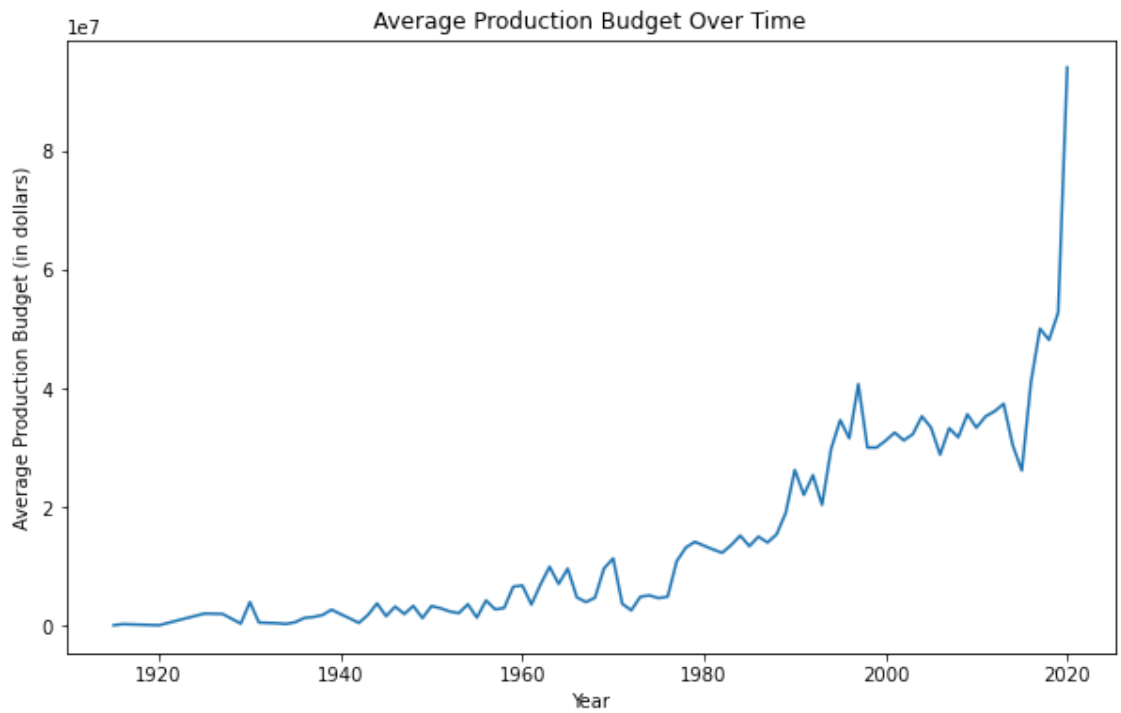
In [46]:
```
print(rt_reviews_cleaned.columns)
print(rt_reviews_cleaned.dtypes)
```

```
Index(['id', 'review', 'rating', 'fresh', 'critic', 'top_critic', 'publis
her',
       'date'],
      dtype='object')
id             int64
review        object
rating        object
fresh         object
critic        object
top_critic     int64
publisher     object
date          object
dtype: object
```

Trends in Production Budget Over Time:

In [49]:

```
# VIZ 3: Trends in Production Budget Over Time
plt.figure(figsize=(10, 6))
tn_movie_budgets_cleaned['year'] = pd.to_datetime(tn_movie_budgets_cleaned
budget_trend = tn_movie_budgets_cleaned.groupby('year')['production_budget
sns.lineplot(x=budget_trend.index, y=budget_trend.values)
plt.title('Average Production Budget Over Time')
plt.xlabel('Year')
plt.ylabel('Average Production Budget (in dollars)')
plt.show()
```



The average production budget has been increasing over the years, indicating growing investments in film production. Recommendation: Plan for higher production budgets in the future to stay competitive and meet audience expectations for high-quality films.

Regression Analysis

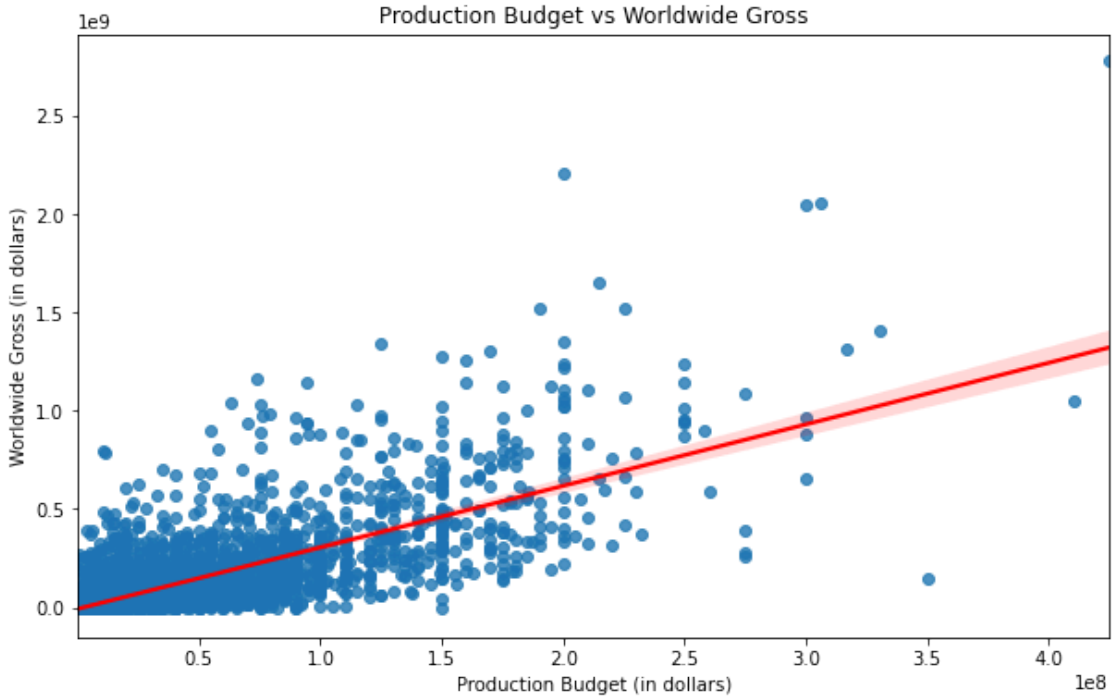
In [76]:

```
import scipy.stats as stats
import statsmodels.api as sm
# Predicting Worldwide Gross based on Production Budget
X = tn_movie_budgets_cleaned['production_budget']
y = tn_movie_budgets_cleaned['worldwide_gross']
X = sm.add_constant(X)  # Adds a constant term to the predictor
model = sm.OLS(y, X).fit()
predictions = model.predict(X)
print(model.summary())
```

```
                            OLS Regression Results
==============================================================================
=====
Dep. Variable:        worldwide_gross   R-squared:
0.560
Model:                            OLS   Adj. R-squared:
0.560
Method:                 Least Squares   F-statistic:
7355.
Date:                Tue, 06 Aug 2024   Prob (F-statistic):
0.00
Time:                        10:25:48   Log-Likelihood:            -1.155
7e+05
No. Observations:                5782   AIC:                         2.31
1e+05
Df Residuals:                    5780   BIC:                         2.31
1e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
============
                        coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
------------
const             -7.286e+06   1.91e+06     -3.813      0.000    -1.1e+07
-3.54e+06
production_budget     3.1269      0.036     85.763      0.000       3.055
3.198
==============================================================================
=====
Omnibus:                     4232.022   Durbin-Watson:
1.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           17239
8.262
Skew:                           3.053   Prob(JB):
0.00
Kurtosis:                      29.044   Cond. No.                     6.5
7e+07
==============================================================================
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is co
rrectly specified.
[2] The condition number is large, 6.57e+07. This might indicate that the
re are
strong multicollinearity or other numerical problems.
```

In [77]:
```
# VIZ 4:Plotting Regression Results
plt.figure(figsize=(10, 6))
sns.regplot(x='production_budget', y='worldwide_gross', data=tn_movie_budg
plt.title('Production Budget vs Worldwide Gross')
plt.xlabel('Production Budget (in dollars)')
plt.ylabel('Worldwide Gross (in dollars)')
plt.show()
```



Insights:

The production budget has a strong positive impact on the worldwide gross, implying that higher investments in production typically result in higher revenues. Studios should consider allocating more resources to production budgets for potentially higher returns. Since the R-squared is 0.560, other factors not included in the model explain 44% of the variance in worldwide gross.

General Recommendations

- Optimal Budget: Invest in movies within the optimal budget range that yields the highest return on investment.
- Genre Focus: Prioritize genres that consistently perform well at the box office. This can include genres like action, adventure, or family movies.
- Quality Over Quantity: Focus on improving both audience and critic ratings as higher ratings correlate with higher revenue. .