Final Project Submission

Please fill out:

- Student name: Magdalene Ondimu
- Student pace: part time
- Scheduled project review date/time: 16th February 2024
- Instructor name: William Okomba, Noah Kandie, Samuel G. Mwangi
- Blog post URL:

```
import matplotlib.pyplot as plt
import matplotlib.image as mpimg

# Open the image file
image_path = r"C:\Users\Magda\OneDrive\Documents\Flatiron\
Picture1.jpg"

# Display the image
img = mpimg.imread(image_path)
plt.imshow(img)
plt.axis('off')  # Turn off axis numbers and ticks
plt.show()
```

Microsoft Movie Studio

Author: Magdalene Ondimu

Business Understanding

1. Microsoft is the stakeholder their interest right now is to create a new movie studio. The problem they are facing is they dont know anything about movies hence they have hired me to help them better understand the movie industry.
2. My task is to help them first of all to better understand the movie industry by, exploring what type of films are currently doing the best at the box office. This i can do by doing:

a) A genre analysis that is to analyze the distribution of movies by genre, the objective of this is to understand the prevalence of different genres in the dataset.

b) A ratings analysis that is to analyze the distribution of movie ratings (popularity), and its objective is to understand the audience preferences based on movie ratings.

c) A market analysis which will entail analyzing the global market potential of movies, its main objective is to Understand the total gross revenue of movies over time and across different countries

d) A competitive analysis which will involve comparing the performance of movies produced by existing studios, its objective will be to identify potential gaps in the market and assess the competitive landscape and also analyze the total gross revenue of movies produced by top studios over time.

1. By doing all the above i will be able to give Microsoft(Stakeholder) a better understanding of the movie industry and the kind of films they would want to create, or produce as they immerse themselves into the movie industry.

Data Understanding

1. The data source is from the the following sites which are in a folder called zippedData: Box Office MojoLinks to an external site. IMDBLinks to an external site. Rotten TomatoesLinks to an external site. TheMovieDBLinks to an external site. The NumbersLinks to an external site. These datasets are suitable for the project as they are from movie studios and also critics who have been in the movie industry for long, these are the better sources to learn from and get a feel of the movie industry.
2. The datasets that i used were:

i) cleaned_bom.movie_gross.csv

ii) cleaned_title.basics.csv

iii) cleaned_tmdb.movies.csv

iv) cleaned_tn.movie_budgets.csv Which i concatenated and used the joined dataset for visualization joined_movies.csv. I got the dataframe level summary statistics by using:

a) .info() - provides the information about the characteristics of the dataframes.

b) .describe()- this is usually used to dig into the summary statistics of the dataset, and get a feel for the data each column contains.

1. The limitations of the data is that the data was not current so as to see the effects on social media and the movie industry. That would have been a very interesting objective to look at.

Data Preparation

1. I got all the datasets provided and started by first of all cleaning them by:

a) reading all the csv files provided

b) using .info() - this is used to get information about the characteristics of the dataframes which tells us: The number of columns and rows in the DataFrame The data type of the data each column contains How many values each column contains (NaNs are not counted) The memory footprint of the DataFrame This sort of information about a dataset is called metadata, since it's data about our data.

c) Using .describe() - this is to dig into the summary statistics of the dataset, and get a feel for the data each column contains.This method is very handy, and gives us relevant information such as: a count of the number of values in each column, making it identify columns with missing values The mean and standard deviation of each column The minimum and maximum values found in each column The median (50%) and quartile values (25% & 75%) for each column 2.This helped me in getting to know how my datasets looked like which data I can use for my analysis.

I will give an example of one dataset and the codes i used because this applies to all the datasets that i used.

```
# import libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
%matplotlib inline

# read the csv file will use one as an example
bom = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\Flatiron\
bom.movie_gross.csv")
bom
```

```
                                              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
```

```
...                                                ...          ...
...
3382                                     The Quake       Magn.
6200.0
3383                   Edward II (2018 re-release)           FM
4800.0
3384                                      El Pacto        Sony
2500.0
3385                                      The Swan  Synergetic
2400.0
3386                             An Actor Prepares       Grav.
1700.0

     foreign_gross  year
0        652000000  2010
1        691300000  2010
2        664300000  2010
3        535700000  2010
4        513900000  2010
...            ...   ...
3382           NaN  2018
3383           NaN  2018
3384           NaN  2018
3385           NaN  2018
3386           NaN  2018

[3387 rows x 5 columns]
```

```
# using .info() to get the information about the characteristics of
the dataframe
bom.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
```

As you can see above i get to see:

1. The type of of my dataframe which is a class dataframe.
2. It gives me the range of my data that is it has 3387 entries which range from 0 to 3386.
3. The number of columns in my dataset are 5 columns in number.

4. It goes further to give me how my dataframe looks like. It give me the titles or column names of my dataset, including their counts that is the entries in each column and the datatype of each column.
5. It summarizes for me the datatypes of the columns that is float is in one column, int in one column and object are in three columns which brings a sum total of the columns to be three. In this summary I can be able to see which columns have missing values and how I can handle them when am analyzing the data.The total entries expected for each column is 3387 so i look into that and see how to handle those below that.

```
# Using .describe() to get summary statistics of the dataset.
bom.describe()
```

```
       domestic_gross         year
count    3.359000e+03  3387.000000
mean     2.874585e+07  2013.958075
std      6.698250e+07     2.478141
min      1.000000e+02  2010.000000
25%      1.200000e+05  2012.000000
50%      1.400000e+06  2014.000000
75%      2.790000e+07  2016.000000
max      9.367000e+08  2018.000000
```

This only showed me two columns that could be calculated. I realized i need the column shown on .info() that is foreign gross to also be included in my analysis so i needed to change its datatype so as it could be included in my summary statistics.

```
# Convert the 'foreign_gross' column to string type
bom['foreign_gross'] = bom['foreign_gross'].astype(str)

# Remove commas from the 'foreign_gross' columb
bom['foreign_gross'] = bom['foreign_gross'].str.replace(',', '')

# Convert the 'foreign_gross' column to float
bom['foreign_gross'] = bom['foreign_gross'].astype(float)
bom
```

```
                                             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
...                                          ...         ...
...
```

```
3382                                       The Quake       Magn.
6200.0
3383                    Edward II (2018 re-release)           FM
4800.0
3384                                        El Pacto        Sony
2500.0
3385                                        The Swan  Synergetic
2400.0
3386                               An Actor Prepares       Grav.
1700.0

      foreign_gross  year
0       652000000.0  2010
1       691300000.0  2010
2       664300000.0  2010
3       535700000.0  2010
4       513900000.0  2010
...             ...   ...
3382            NaN  2018
3383            NaN  2018
3384            NaN  2018
3385            NaN  2018
3386            NaN  2018

[3387 rows x 5 columns]
```

```
# so lets check if the foreign_gross could be used in the summary statistics
bom.describe()
```

```
       domestic_gross  foreign_gross         year
count    3.359000e+03   2.037000e+03  3387.000000
mean     2.874585e+07   7.487281e+07  2013.958075
std      6.698250e+07   1.374106e+08     2.478141
min      1.000000e+02   6.000000e+02  2010.000000
25%      1.200000e+05   3.700000e+06  2012.000000
50%      1.400000e+06   1.870000e+07  2014.000000
75%      2.790000e+07   7.490000e+07  2016.000000
max      9.367000e+08   9.605000e+08  2018.000000
```

Voila! thats is done.

I did this to most datasets what needed to be used and changed from one datatype to another I did so. I looked at the missing values and decided for some to use the median to fill in the missing values and for some i used the ffill() method to fill in the missing values. This is so because I did not want to drop any columns or rows as I saw that they were important in the analysis of the project. After doing all these data cleaning on all datasets then I saved the new cleaned data now for proper analysis.

In this stage i used cleaned data and wanted to work with just the datasets that would meet the objective i had for this project. So I was looking at the datasets that would give me the genre, ratings, market and competitive analysis for this particular project. I dropped the datasets that didnt give me the above. Now I started working with the datasets using the above criteria and also dropped some columns that I did not need.

Examples of what I did.

```
# load the cleaned data of the datasets i used.
bom_movie = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\Flatiron\
cleaned_bom.movie_gross.csv")
title_basics = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\
Flatiron\cleaned_title.basics.csv")
tmdb_movies = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\
Flatiron\cleaned_tmdb.movies.csv")
tn_movie = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\Flatiron\
cleaned_tn.movie_budgets.csv")
```

The second step was to concatenate these datasets.

1. Change column names that had the same details to have the same names.
2. Drop the columns that I did not need in the data analysis.
3. Concatenate the datasets.

```
# Concatenate the datasets
joined_movies = pd.concat([bom_movie, tn_movie, tmdb_movies,
title_basics], ignore_index=True)
joined_movies
```

```
                                                 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
...                                               ...    ...
...
181825                            Kuambil Lagi Hatiku    NaN
NaN
181826  Rodolpho Teóphilo - O Legado de um Pioneiro    NaN
NaN
181827                                Dankyavar Danka    NaN
NaN
181828                                        6 Gunn    NaN
```

```
NaN
181829                Chico Albuquerque - Revelações    NaN
NaN

        foreign_gross  year  id production_budget  Unnamed: 0
genre_ids  \
0         652000000.0  2010 NaN                NaN         NaN
NaN
1         691300000.0  2010 NaN                NaN         NaN
NaN
2         664300000.0  2010 NaN                NaN         NaN
NaN
3         535700000.0  2010 NaN                NaN         NaN
NaN
4         513900000.0  2010 NaN                NaN         NaN
NaN
...               ...   ...  ..                ...         ...        ..
.
181825            NaN  2019 NaN                NaN         NaN
NaN
181826            NaN  2015 NaN                NaN         NaN
NaN
181827            NaN  2013 NaN                NaN         NaN
NaN
181828            NaN  2017 NaN                NaN         NaN
NaN
181829            NaN  2013 NaN                NaN         NaN
NaN

       original_language                                original_title
\
0                    NaN                                           NaN

1                    NaN                                           NaN

2                    NaN                                           NaN

3                    NaN                                           NaN

4                    NaN                                           NaN

...                  ...                                           ...

181825               NaN                           Kuambil Lagi Hatiku

181826               NaN  Rodolpho Teóphilo - O Legado de um Pioneiro

181827               NaN                               Dankyavar Danka

181828               NaN                                       6 Gunn
```

```
181829               NaN               Chico Albuquerque - Revelações


        popularity  vote_average  vote_count     tconst
runtime_minutes  \
0              NaN           NaN         NaN        NaN
NaN
1              NaN           NaN         NaN        NaN
NaN
2              NaN           NaN         NaN        NaN
NaN
3              NaN           NaN         NaN        NaN
NaN
4              NaN           NaN         NaN        NaN
NaN
...            ...           ...         ...        ...                .
..
181825         NaN           NaN         NaN  tt9916538
123.000000
181826         NaN           NaN         NaN  tt9916622
120.666667
181827         NaN           NaN         NaN  tt9916706
118.333333
181828         NaN           NaN         NaN  tt9916730
116.000000
181829         NaN           NaN         NaN  tt9916754
116.000000

             genres
0               NaN
1               NaN
2               NaN
3               NaN
4               NaN
...             ...
181825        Drama
181826  Documentary
181827       Comedy
181828          NaN
181829  Documentary

[181830 rows x 17 columns]

# Drop the columns
joined_movies.drop(columns=['id', 'production_budget', 'genre_ids',
'original_language', 'original_title','vote_average',
                            'vote_count','tconst','runtime_minutes'],
inplace=True)
joined_movies
```

```
                                                 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
...                                              ...    ...
...
181825                           Kuambil Lagi Hatiku    NaN
NaN
181826  Rodolpho Teóphilo - O Legado de um Pioneiro    NaN
NaN
181827                               Dankyavar Danka    NaN
NaN
181828                                       6 Gunn    NaN
NaN
181829               Chico Albuquerque - Revelações    NaN
NaN

        foreign_gross  year  Unnamed: 0  popularity       genres
0         652000000.0  2010         NaN         NaN          NaN
1         691300000.0  2010         NaN         NaN          NaN
2         664300000.0  2010         NaN         NaN          NaN
3         535700000.0  2010         NaN         NaN          NaN
4         513900000.0  2010         NaN         NaN          NaN
...               ...   ...         ...         ...          ...
181825            NaN  2019         NaN         NaN        Drama
181826            NaN  2015         NaN         NaN  Documentary
181827            NaN  2013         NaN         NaN       Comedy
181828            NaN  2017         NaN         NaN          NaN
181829            NaN  2013         NaN         NaN  Documentary

[181830 rows x 8 columns]
```

After concatenating the four datasets and dropping the columns that will not be used in my analysis now i went further on to start working on my datasets so as to get information needed by the stakeholder.

Data Analysis

The stakeholder should put this in consideration:

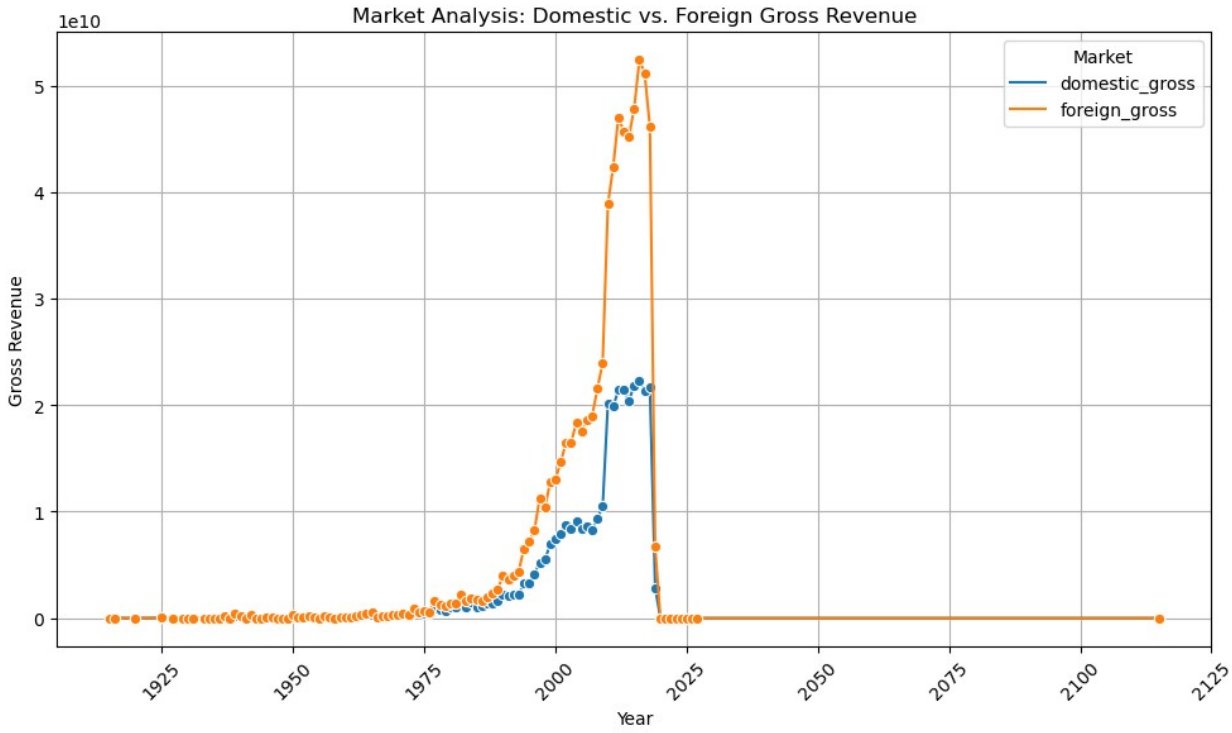
1. The market analysis showed that the foreign market has evolved overtime. The foreign market has been dominant in contributing to the overall revenue in different years.The

foreign market embraced the blockbuster movies than the domestic market. The stakeholder can embark on having an appeal to the foreign market as this will boost their revenue tremendously.

2. In the genre analysis i looked at the top 10 movie genres. In this the stakeholder can start producing more movies according to the genres as they will appeal to audience more.
3. The total gross revenue by the top 10 studios is also something the stakeholder should consider as this will help them know the studios which they can partner with to get the maximum of their return.

Visualization

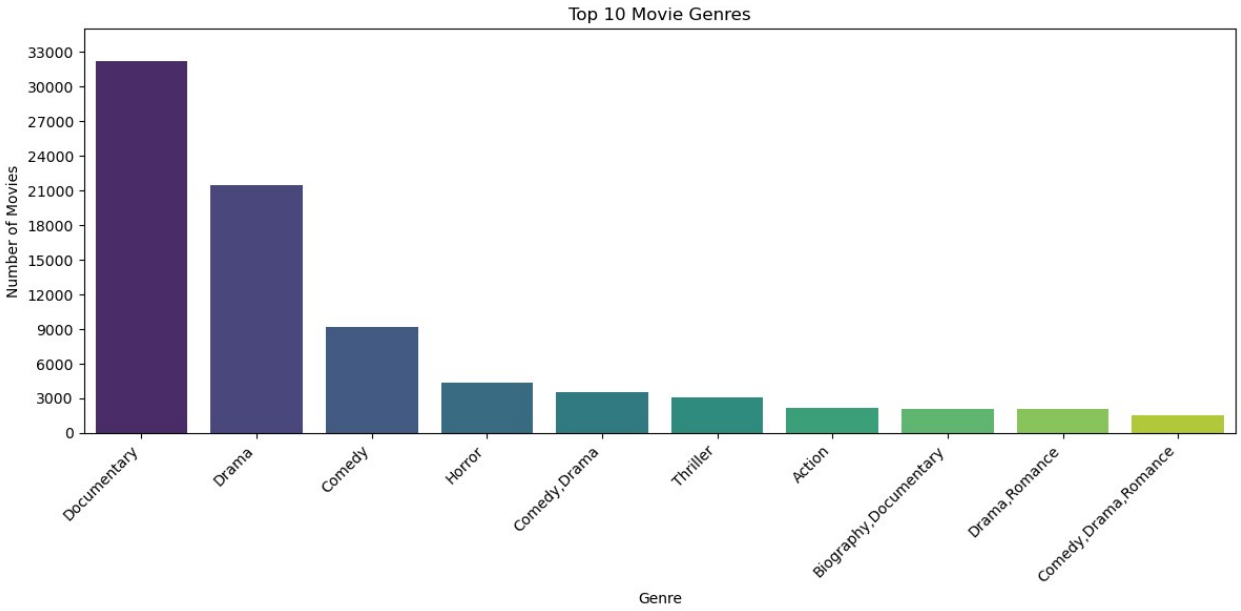
```
# Load the joined data
movies_data = pd.read_csv(r"C:\Users\Magda\OneDrive\Documents\
Flatiron\joined_movies.csv", low_memory=False)

# Market analysis on comparing the foreign and domestic markets.
# Convert 'title_year' column to datetime
#movies_data['year'] = pd.to_datetime(movies_data['year'],
format='%Y')

# Group by year and calculate total domestic and foreign gross revenue
for each year
market_analysis = movies_data.groupby(movies_data['year'].dt.year)
[['domestic_gross', 'foreign_gross']].sum().reset_index()

# Melt the DataFrame to long format for visualization
market_analysis_melted = market_analysis.melt(id_vars=['year'],
var_name='Market', value_name='Gross Revenue')

# Plotting
plt.figure(figsize=(10, 6))
sns.lineplot(data=market_analysis_melted, x='year', y='Gross Revenue',
hue='Market', marker='o')
plt.title('Market Analysis: Domestic vs. Foreign Gross Revenue')
plt.xlabel('Year')
plt.ylabel('Gross Revenue')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



The above trend shows that the foreign market grossed more than the domestic market. This gives the stakeholder insight on where they can focus more of how they can improve to enable them to have the domestic market watch more movies in studios.

```
# Genre analysis on which genres the new movie studio can focus on
this shows the top 10.
# Split genres into separate rows
genres = movies_data['genres'].str.split('|',
expand=True).stack().reset_index(level=1, drop=True)
movies_data_split = movies_data.drop('genres',
axis=1).join(genres.rename('genre'))

# Get top 10 genres by count
top_10_genres =
movies_data_split['genre'].value_counts().head(10).index

# Filter DataFrame for top 10 genres
movies_data_top_genres =
movies_data_split[movies_data_split['genre'].isin(top_10_genres)]

# Plotting
plt.figure(figsize=(12, 6))
sns.countplot(data=movies_data_top_genres, x='genre',
palette='viridis', order=top_10_genres)
plt.title('Top 10 Movie Genres')
plt.xlabel('Genre')
plt.ylabel('Number of Movies')
```

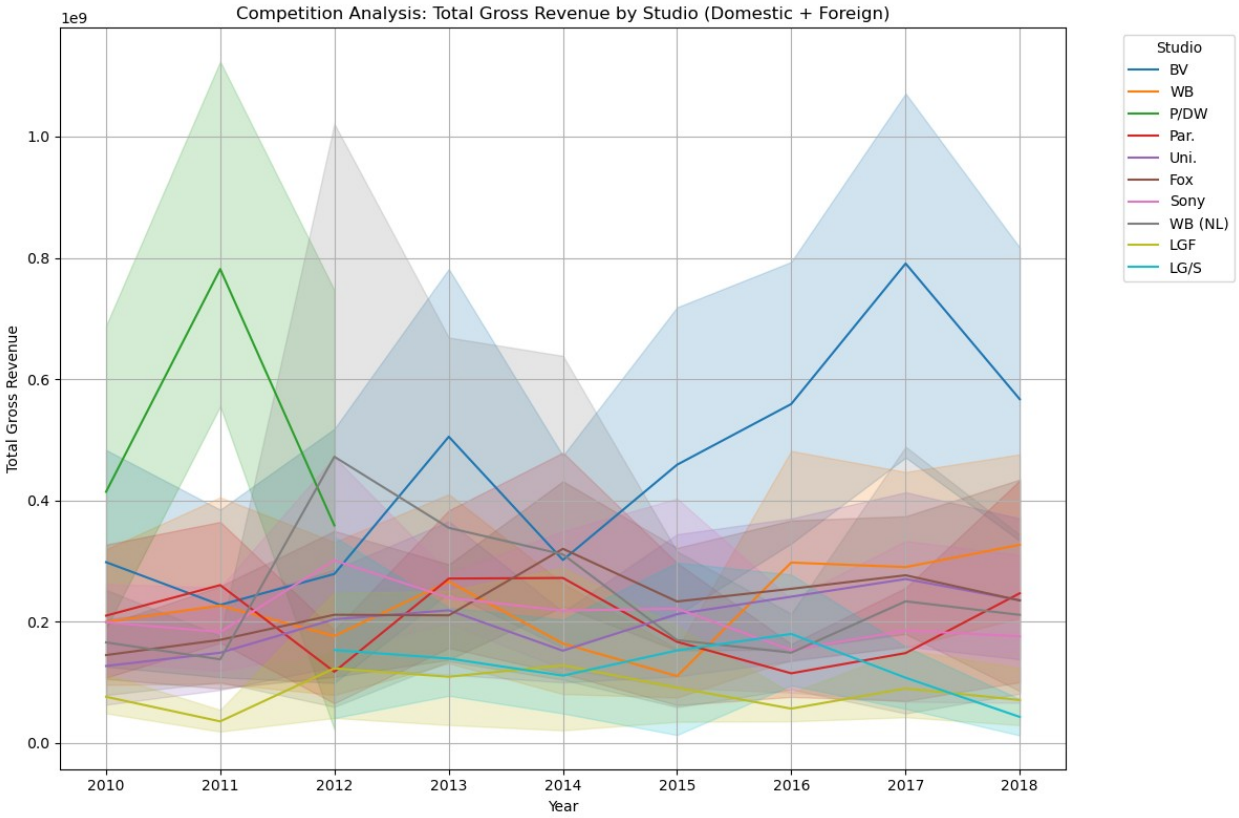
```
plt.xticks(rotation=45, ha='right')
plt.yticks(range(0, 35001, 3000))  # Adjust y-axis ticks up to 35000
plt.ylim(0, 35000)  # Set y-axis upper limit to 35000
plt.tight_layout()
plt.show()
```



The recommendation to the stakeholder is to invest more in documentaries, drama, comedy, horror, comedy-drama, thriller, action, biography-documentary, drama-romance and comedy-drama-romance as these are the most watched.

```
# Competition analysis between studios these are the top 10 studios
which grossed more than the rest.
# Calculate total gross revenue by summing domestic and foreign gross
movies_data['total_gross'] = movies_data['domestic_gross'] +
movies_data['foreign_gross']

# Group by studio and calculate total gross revenue for each studio
studio_total_gross = movies_data.groupby('studio')
['total_gross'].sum().sort_values(ascending=False)

# Select the top 10 studios
top_10_studios = studio_total_gross.head(10).index

# Filter DataFrame for top 10 studios
movies_data_top_studios =
movies_data[movies_data['studio'].isin(top_10_studios)]

# Plotting
plt.figure(figsize=(12, 8))
sns.lineplot(data=movies_data_top_studios, x='year', y='total_gross',
```

```
hue='studio')
plt.title('Competition Analysis: Total Gross Revenue by Studio
(Domestic + Foreign)')
plt.xlabel('Year')
plt.ylabel('Total Gross Revenue')
plt.legend(title='Studio', bbox_to_anchor=(1.05, 1), loc='upper left')
plt.grid(True)
plt.tight_layout()
plt.show()
```



This shows the stakeholder which movie studios they can partner with so as to produce movies which are most watched and gross above the rest.