Phase 2 Project

A company decided to create a new movie studio and needs help deciding what type of movies to make. I am also going to make the assumption that this company is based in North America and its target market is an English-speaking audience.

Initial datasets are from:

- Box Office Mojo
- IMDB
- Rotten Tomatoes
- TheMovieDB
- The Numbers

Additional data was scrapped from:

- Box Office Mojo, to obtain data from the past 2 years.

Importing Libraries

In [ ]:

```
import sqlite3
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

Which genres are highly rated?

I started with the Rotten Tomatoes datasets to explore genres and use Fresh (60% and higher) vs. Rotten as my metric.

In [ ]:

```
rt = pd.read_csv("data/rt.csv")
rt
```

In [ ]:

```
#Explode genre column
rt["genre"] = rt["genre"].str.split("|")
rt = rt.explode("genre")
rt["genre"].value_counts()
```

After looking at the value counts, I wanted to filter what genres would potentially skew the data and would have to be removed.

- Classics are much older films and I would not consider them relevant to this analysis.
- Is Television the same as streaming in this data set?
- What is Special Interest, and is it too narrow?
- The genres Cult Movies, Gay and Lesbian, Anime and Manga had under 100 and I consider them to not have enough information when looking at genres.

In [ ]:

```
#What is the Television genre
rt["theater_date"] = pd.to_datetime(rt["theater_date"])
rt.sort_values("theater_date", ascending=False).head(15)
rt_tv = rt[rt["genre"] == "Television"]
rt_tv.sort_values("theater_date", ascending=False)
```

In [ ]:

```
#What is the Special Interest genre
rt_si = rt[rt["genre"] == "Special Interest"]
rt_si.sample(15)
```

In [ ]:

```
rt = rt.drop(rt[rt.genre.isin(["Cult Movies", "Gay and Lesbian", "Anime and Manga", "Classics", "Television", "Special Interest"])].index)
```

Looking at Fresh vs. Rotten ratio.

In [ ]:

```
rt_freshvsrotten = rt.groupby(["genre", "fresh"]).count()["id"]
rt_freshvsrotten = rt_freshvsrotten.reset_index().pivot(columns="fresh", index="genre", values="id")
rt_freshvsrotten
```

In [ ]:

```
rt_freshvsrotten["fresh / rotten"] = rt_freshvsrotten["fresh"] / rt_freshvsrotten["rotten"]
rt_freshvsrotten = rt_freshvsrotten.sort_values("fresh / rotten", ascending = False)
```

In [ ]:

```
sns.set(style="darkgrid")
ax = sns.barplot(data=rt_freshvsrotten, x="fresh / rotten", y=rt_freshvsrotten.index, orient="h")
ax.set_title("Highest Rated Genres on Rotten Tomatoes")
ax.set(xlabel="Fresh vs. Rotten Ratio", ylabel="Genres");
```

From this, the top 5 genres that are highly rated are: Documentary, Art House and International, Sports and Fitness, Animation, and Drama. Noting that Romance is in sixth place, if International being attached to Art House is a concern later in this analysis.

Which of these genres make the most profit on average?

For this, the first step was to sort the data from IMDB. To follow focusing on the highly rated, I filtered this dataset to return the average rating was above 6.0 similar to Rotten Tomatoes' 60% and higher Fresh rating and number of votes had to be above 100 to remove outliers. I also focused on world revenue rather than domestic revenue due to English-speaking being a predominant international language.

In [ ]:

```
conn = sqlite3.connect("data/im.db")
```

In [ ]:
```
imdb = pd.read_sql("""
SELECT persons.primary_name as 'Name', persons.primary_profession as 'Profession', movie_basics.primary_title as 'Title', movie_basics.genres as 'Genres', movie_ratings
FROM persons
INNER JOIN directors
ON persons.person_id = directors.person_id
INNER JOIN writers
ON directors.person_id = writers.person_id
INNER JOIN movie_basics
ON directors.movie_id = movie_basics.movie_id
INNER JOIN movie_ratings
ON movie_basics.movie_id = movie_ratings.movie_id
WHERE averagerating > 6.0 and numvotes > 100
""", conn)
imdb
```

In [ ]:
```
imdb["Genres"] = imdb["Genres"].str.split(",")
imdb = imdb.explode("Genres")
imdb["Genres"].value_counts()
```

In [ ]:
```
imdb.dropna(subset=["Genres"], inplace=True)
```

Due to how IMDB categorizes its movies, there was neither a "Art House" nor "International" genre to reference, so I did have to alter my "top 5" to include now Romance.

In [ ]:
```
imdb = imdb[imdb["Genres"].str.contains("Documentary|Sport|Animation|Drama|Romance")]
imdb
```

Merginging this data with The Numbers CSV for product budgets and worldwide gross.

In [ ]:
```
tn_df = pd.read_csv("data/tn.movie_budgets.csv")
tn_df
```

In [ ]:
```
merge = imdb.merge(tn_df, left_on='Title', right_on='movie',how='inner')
[['Name','Title','Genres', 'Production Budget', 'Domestic Gross', 'Foreign Gross', 'Average Rating', 'NumVotes']]
```

In [ ]:
```
merge["production_budget"] = merge["production_budget"].replace('[\$,]', '', regex=True).astype(int)
merge["domestic_gross"] = merge["domestic_gross"].replace("[\$,]", "", regex=True).astype(int)
merge["worldwide_gross"] = merge["worldwide_gross"].replace("[\$,]", "", regex=True).astype(int)
merge["avg_world"] = merge["worldwide_gross"] / merge["production_budget"]
merge
```

In [ ]:
```
genre_wprofit = merge.groupby("Genres")["avg_world"].mean().sort_values(ascending=False)
genre_wprofit = genre_wprofit.reset_index()
genre_wprofit
```

In [ ]:
```
fig, ax = plt.subplots(figsize = (10,5))
plt.bar(data=genre_wprofit, x=genre_wprofit["Genres"], height=genre_wprofit["avg_world"])
plt.xticks(range(len(genre_wprofit["Genres"])), genre_wprofit["Genres"])
plt.xlabel("Top Genres Selected")
plt.ylabel("Average Profit Ratio")
plt.title("Average of Worldwide Gross / Production Budget");
```

From the new top 5 genres, Animation has the highest average profit.

Which director and writer are highly rated for Animation?

Going back to the starting IMDB data and now filtering the genre to Animation only.

In [ ]:
```
#Exploring Animation
animation = imdb[imdb["Genres"] == "Animation"]
animation
```

In [ ]:
```
animation["Profession"] = animation["Profession"].str.split(",")
animation = animation.explode("Profession")
```

In [ ]:
```
directors = animation[animation["Profession"] == "director"]
directors = directors.drop_duplicates(subset=["Title"])
directors
```

Now that this data is filtered by directors, which directors on average have the highest ratings?

In [ ]:
```
avg_drating = directors.groupby("Name")["Average Rating"].mean().sort_values(ascending=False)
avg_drating.head(15)
```

Looking into these director's IMDB pages, Mert Gökalp and Harry Baweja are international directors and although their films are English subtitled, that would be additional step for this company to take when starting up. Therefore, the next best pick is Rodney Rothman, who worked on the Marvel Spiderverse films. Third is Pete Docter, who director on a few Pixar films.

Similar to directors, who are the highest rated writers in Animation?

In [ ]:
```
writers = animation[animation["Profession"] == "writer"]
writers = writers.drop_duplicates(subset=["Title"])
avg_wrating = writers.groupby("Name")["Average Rating"].mean().sort_values(ascending=False)
avg_wrating.head(15)
```

A similar situation to directors regarding Mert Gökalp. Rodney Rothman is second again here. Adrian Molina is also a good pick as he was the co-director for Disney's Coco.

What is the best month to make a profit?

With the available data, I also wanted to look for which month on average made the most profit as insight on when to release this animated movie.

In [ ]:
```
merge["release_date"]= pd.to_datetime(merge["release_date"])
```

In [ ]:
```
animation_month = merge[merge["Genres"] == "Animation"]
animation_month = animation_month.drop_duplicates(subset=["id"])
animation_month["month"] = animation_month["release_date"].dt.month
chart = animation_month.groupby("month")["worldwide_gross"].mean()
chart = chart.reset_index()
chart
```

In [ ]:

```
fig, ax = plt.subplots()
sns.set(style="darkgrid")
linechart = sns.lineplot(data=chart, x="month", y="worldwide_gross")
ax.set_xticks([1,2,3,4,5,6,7,8,9,10,11,12],["Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec"], rotation=25)
ax.set(xlabel="Month", ylabel="In Trillions, AVG World Gross")
ax.set_title("Average Worldwide Gross per Month");
```

From this data, it appears that March, followed by November, are the best months to release an animated movie. It should be noted there is no data for January, which could imply that a November release is meant to carry throughout the winter holidays.

What is the current trend of animation?

Since most of the initial data was from before 2020, I thought it would be important to see how the movie industry is doing after COVID-19. This scrapped data focuses on revenue only.

In [ ]:

```
bom = pd.read_csv("data/bom_scrapy.csv", low_memory=False)
bom
```

In [ ]:

```
bom.drop_duplicates()
```

In [ ]:

```
#Drop films that are too new, they will skew the data
bom = bom.drop(bom[bom["release_days"] <= 60].index)
```

In [ ]:

```
#Explode genre column
bom["genres"] = bom["genres"].str.split(",")
bom = bom.explode("genres")
```

In [ ]:

```
bom["genres"].value_counts()
```

In [ ]:

```
bom = bom[bom["genres"].str.contains("Documentary|Sport|Animation|Drama|Romance")]
```

In [ ]:

```
#Data cleaning
bom["domestic_revenue"].fillna(1, inplace=True)
bom["world_revenue"].fillna(1, inplace=True)
bom["opening_revenue"].fillna(1, inplace=True)
bom["domestic_revenue"] = bom["domestic_revenue"].replace("[\$,]", "", regex=True).astype(int)
bom["world_revenue"] = bom["world_revenue"].replace("[\$,]", "", regex=True).astype(int)
bom["opening_revenue"] = bom["opening_revenue"].replace('[\$,]', '', regex=True).astype(int)
```

In [ ]:

```
genre_wprofit = bom.groupby("genres")["world_revenue"].mean().sort_values(ascending=False)
genre_wprofit = genre_wprofit.reset_index()
genre_wprofit
```

In [ ]:

```
fig, ax = plt.subplots(figsize = (10,5))
plt.bar(data=genre_wprofit, x=genre_wprofit["genres"], height=genre_wprofit["world_revenue"])
plt.xticks(range(len(genre_wprofit["genres"])), genre_wprofit["genres"])
plt.xlabel("Top Genres Selected")
plt.ylabel("In Trillions, AVG Profit")
plt.title("Current: Average of Worldwide Gross");
```

According to the scrapped data from Box Office Mojo, animation as a genre compared to the top 5 previously considered is doing well.

In [ ]:

```
conn.close()
```

Conclusion

According to the available data, amongst the highly rated genres, Animation generates the most world revenue on average. For creatives, Rodney Rothman and Pete Docter good picks for directors. For writers, Rodney Rothman is again a good pick, as well as Adrian Molina. The best month for profit in releasing an animated movie is November, followed by March. In regards to more current data, animation is still doing the best profit wise compared the highly rated genres from earlier.