Hollywood by the Numbers: Trends from The Movies Dataset

Presented by: Group 5iveCoders

Starring...

Ifrah

The film is never the same the second or ..the hundredth time

Manny

Mercury is the smallest planet in the Solar System

Tin

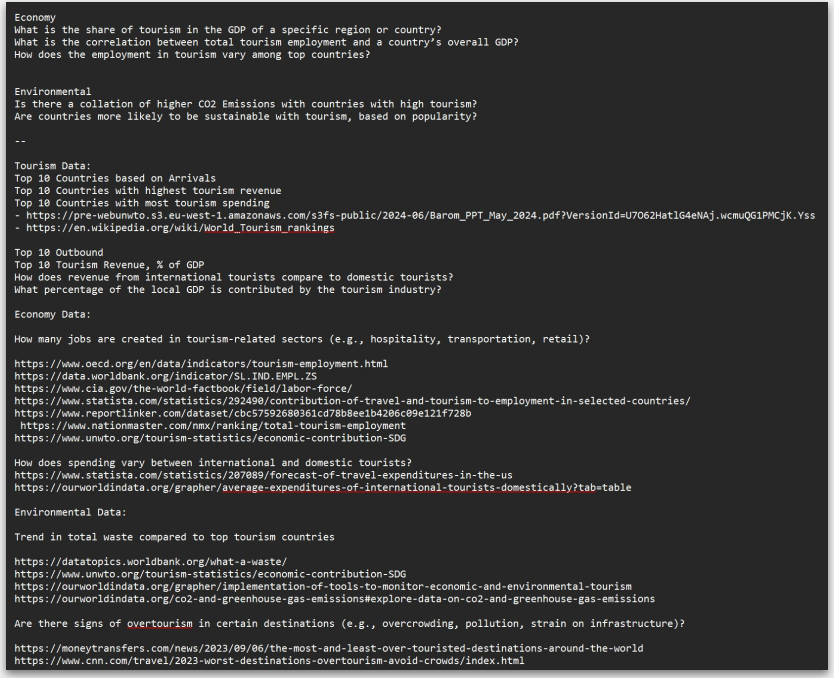
Hot Take:
Antitrust is a good movie!

Trouble in Paradise

We originally scoped our project to analyze the impact of tourism on local economies and environments.

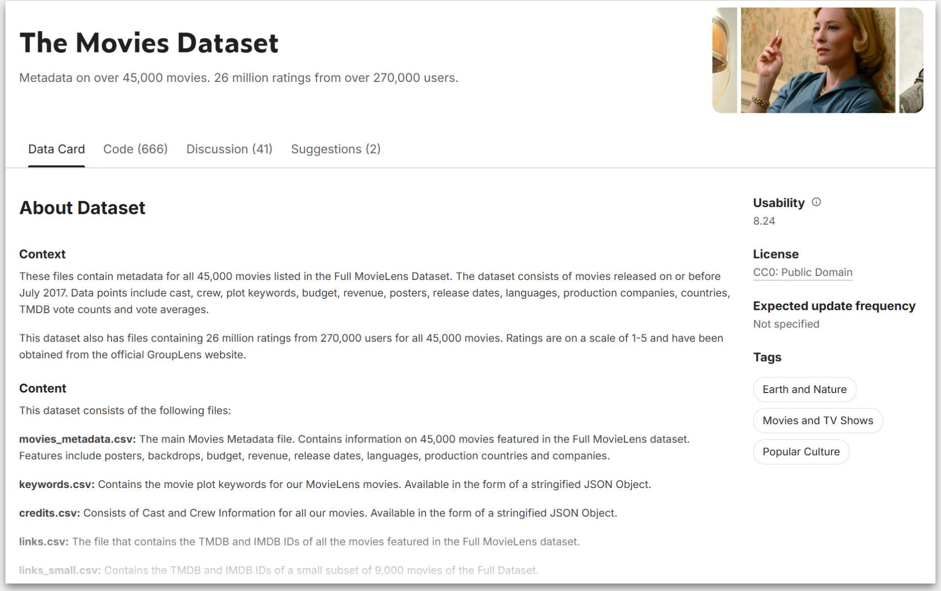
We ran into issues:

- Multiple datasets that required combining
- No single large dataset
- Data entry would have been required for PDF based reports
- Limited Scope on Data/Reporting



```
Economy
What is the share of tourism in the GDP of a specific region or country?
What is the correlation between total tourism employment and a country’s overall GDP?
How does the employment in tourism vary among top countries?


Environmental
Is there a collation of higher CO2 Emissions with countries with high tourism?
Are countries more likely to be sustainable with tourism, based on popularity?

--

Tourism Data:
Top 10 Countries based on Arrivals
Top 10 Countries with highest tourism revenue
Top 10 Countries with most tourism spending
- https://pre-webunwto.s3.eu-west-1.amazonaws.com/s3fs-public/2024-06/Barom_PPT_May_2024.pdf?VersionId=U7O62HatlG4eNAj.wcmuQG1PMCjK.Yss
- https://en.wikipedia.org/wiki/World_Tourism_rankings

Top 10 Outbound
Top 10 Tourism Revenue, % of GDP
How does revenue from international tourists compare to domestic tourists?
What percentage of the local GDP is contributed by the tourism industry?

Economy Data:

How many jobs are created in tourism-related sectors (e.g., hospitality, transportation, retail)?

https://www.oecd.org/en/data/indicators/tourism-employment.html
https://data.worldbank.org/indicator/SL.IND.EMPL.ZS
https://www.cia.gov/the-world-factbook/field/labor-force/
https://www.statista.com/statistics/292490/contribution-of-travel-and-tourism-to-employment-in-selected-countries/
https://www.reportlinker.com/dataset/cbc57592680361cd78b8ee1b4206c09e121f728b
 https://www.nationmaster.com/nmx/ranking/total-tourism-employment
https://www.unwto.org/tourism-statistics/economic-contribution-SDG

How does spending vary between international and domestic tourists?
https://www.statista.com/statistics/207089/forecast-of-travel-expenditures-in-the-us
https://ourworldindata.org/grapher/average-expenditures-of-international-tourists-domestically?tab=table

Environmental Data:

Trend in total waste compared to top tourism countries

https://datatopics.worldbank.org/what-a-waste/
https://www.unwto.org/tourism-statistics/economic-contribution-SDG
https://ourworldindata.org/grapher/implementation-of-tools-to-monitor-economic-and-environmental-tourism
https://ourworldindata.org/co2-and-greenhouse-gas-emissions#explore-data-on-co2-and-greenhouse-gas-emissions

Are there signs of overtourism in certain destinations (e.g., overcrowding, pollution, strain on infrastructure)?

https://moneytransfers.com/news/2023/09/06/the-most-and-least-over-touristed-destinations-around-the-world
https://www.cnn.com/travel/2023-worst-destinations-overtourism-avoid-crowds/index.html
```

The Movies Dataset



The Movies Dataset

Metadata on over 45,000 movies. 26 million ratings from over 270,000 users.

About Dataset

Context

These files contain metadata for all 45,000 movies listed in the Full MovieLens Dataset. The dataset consists of movies released on or before July 2017. Data points include cast, crew, plot keywords, budget, revenue, posters, release dates, languages, production companies, countries, TMDB vote counts and vote averages.

This dataset also has files containing 26 million ratings from 270,000 users for all 45,000 movies. Ratings are on a scale of 1-5 and have been obtained from the official GroupLens website.

Content

This dataset consists of the following files:

movies_metadata.csv: The main Movies Metadata file. Contains information on 45,000 movies featured in the Full MovieLens dataset. Features include posters, backdrops, budget, revenue, release dates, languages, production countries and companies.

keywords.csv: Contains the movie plot keywords for our MovieLens movies. Available in the form of a stringified JSON Object.

credits.csv: Consists of Cast and Crew Information for all our movies. Available in the form of a stringified JSON Object.

links.csv: The file that contains the TMDB and IMDB IDs of all the movies featured in the Full MovieLens dataset.

links_small.csv: Contains the TMDB and IMDB IDs of a small subset of 9,000 movies of the Full Dataset.

Usability ⓘ
8.24

License
CC0: Public Domain

Expected update frequency
Not specified

Tags

Earth and Nature

Movies and TV Shows

Popular Culture

After evaluating our original scope, we decided to pivot and found the The Movies Dataset

Factors that enable our decision:

- Single Dataset with multiple CSVs
  - 45,000 Movies
  - 26 Million Ratings
- No Data Entry Required
- OMDb and TMDb APIs to support data
- A lot of opportunity for deeper analysis

Discovering Trends with the Movies Dataset

1 **Most Common Genres**

In the last decade

2 **Directors**

With the highest IMDb rating

3 **Movie Runtimes**

Are they getting longer or shorter?

4 **Actors**

Who appears the most frequent in high-rated movies?

5 **Movie Budgets**

Does more money mean better movies?

6 **Country Origin**

Which countries are producing the most films?

7 **Genre Revenue**

Which genre creates the most revenue?

8 **Genre Profit**

Genre's revenue correlate to its profitability?

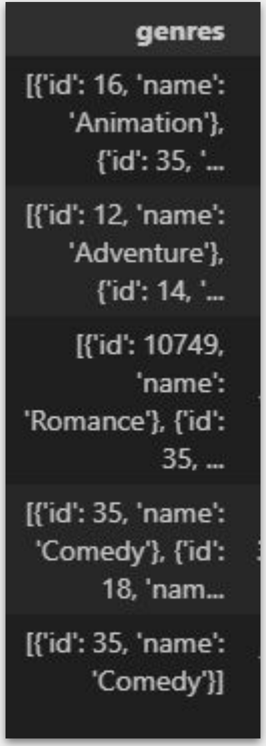
Question 1:
What are the most common movie genres released over the years?

The dataset we used, metadata_movies.csv included movies released as far back as 1990. To simplify the visualization, we focused on movies from the last decades.

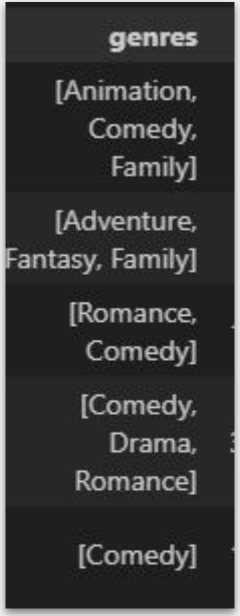
The genres column was originally in dictionary format, so we converted it into a list of genre names for easier processing. The transformation applied was:

```
cleaned_movies_df["genres"] =
cleaned_movies_df["genres"].apply(lambda x: x
if isinstance(x, list) else [])
```

```
cleaned_movies_df["genres"]=cleaned_movies_df["
genres"].apply(lambda x: [i["name"] for i in x]
if isinstance(x, list) else [])
```



genres
[{'id': 16, 'name': 'Animation'}, {'id': 35, '...
[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...
[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...
[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...
[{'id': 35, 'name': 'Comedy'}]



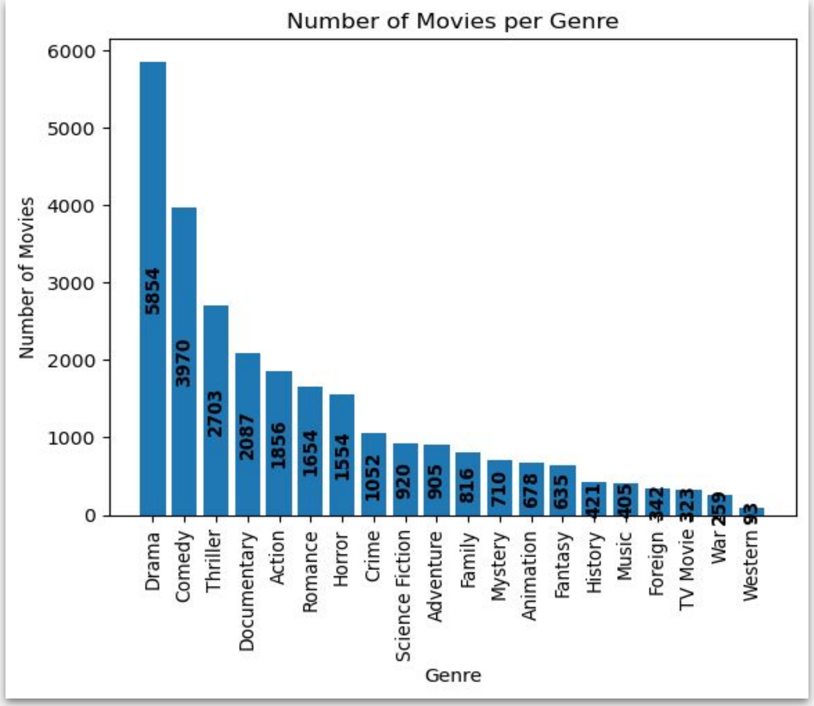
genres
[Animation, Comedy, Family]
[Adventure, Fantasy, Family]
[Romance, Comedy]
[Comedy, Drama, Romance]
[Comedy]

Question 1:
What are the most common movie genres released over the years? (cont.)

Flattened the genres list to count how many movies were released per genre over the years.

```
# Flatten the 'genres' list and count occurrences

genre_counts =
last_decade_movies_df['genres'].explode().value_counts()
# Create a DataFrame with the genre counts
genre_counts_df =
pd.DataFrame(genre_counts).reset_index()
genre_counts_df.columns = ['genre', 'count']
genre_counts_df.head(20)
```

The bar graph shows Drama and Comedy as the most released genres in the last decade.

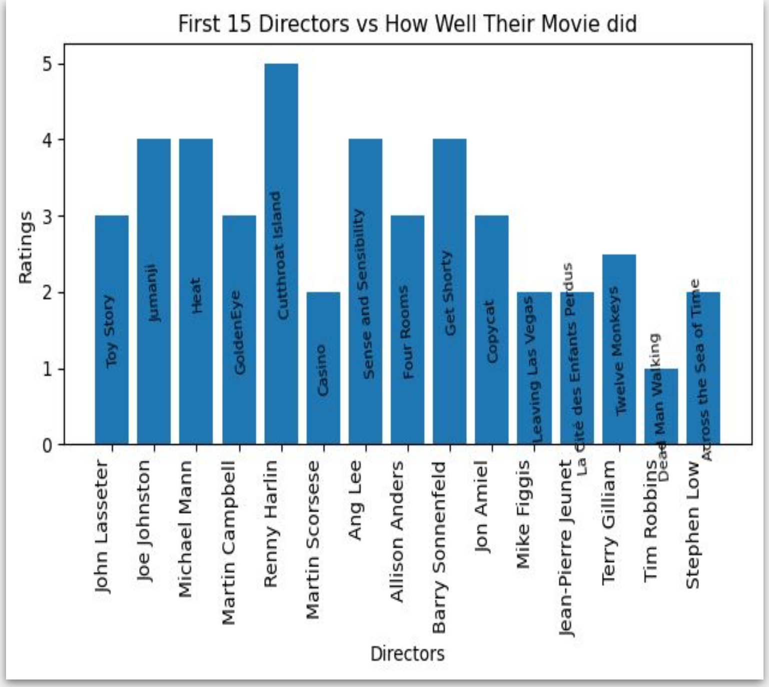


Question 2:
Which directors have the highest average IMDb ratings across their films?

Imported data from various .csv files(credits, ratings, movies.metadata) to merge into one Final_rating_df,DataFrame, to get the first 15 Directors and the imdb_rating for their movies.

Shortcoming:

Sorting ratings in ascending order gave us, the 5.0 ratings of all directors, hence no disparity to go off by thus only going by the first 15_Directors.



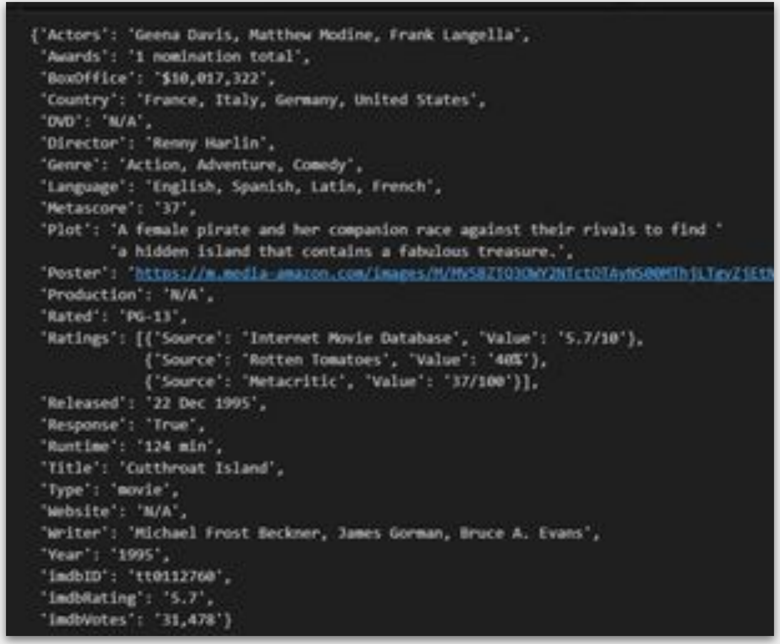
Question 2:
Which directors have the highest average IMDb ratings across their films? (cont).

To get more information on the movie with the highest rating, i did an API search:

```
api_key= "&apikey=" + "99b7bf4e"
url="https://www.omdbapi.com/?t="
response=requests.get(url +
first_rating_df.iloc[4,3] + api_key)
data=response.json()
pprint(data)
```

Hence knowing that:

The Movie Cutthroat Island, was released in 22 Dec 1995, got 1 nomination total and had a Box Office total of: $10,017,322.



```
{'Actors': 'Geena Davis, Matthew Modine, Frank Langella',
 'Awards': '1 nomination total',
 'BoxOffice': '$10,017,322',
 'Country': 'France, Italy, Germany, United States',
 'DVD': 'N/A',
 'Director': 'Renny Harlin',
 'Genre': 'Action, Adventure, Comedy',
 'Language': 'English, Spanish, Latin, French',
 'Metascore': '37',
 'Plot': 'A female pirate and her companion race against their rivals to find '
         'a hidden island that contains a fabulous treasure.',
 'Poster': 'https://m.media-amazon.com/images/M/MV5BZTQ3OWY2NTctOTAyMS00MThjLTgyZjEtN
 'Production': 'N/A',
 'Rated': 'PG-13',
 'Ratings': [{'Source': 'Internet Movie Database', 'Value': '5.7/10'},
             {'Source': 'Rotten Tomatoes', 'Value': '40%'},
             {'Source': 'Metacritic', 'Value': '37/100'}],
 'Released': '22 Dec 1995',
 'Response': 'True',
 'Runtime': '124 min',
 'Title': 'Cutthroat Island',
 'Type': 'movie',
 'Website': 'N/A',
 'Writer': 'Michael Frost Beckner, James Gorman, Bruce A. Evans',
 'Year': '1995',
 'imdbID': 'tt0112760',
 'imdbRating': '5.7',
 'imdbVotes': '31,478'}
```

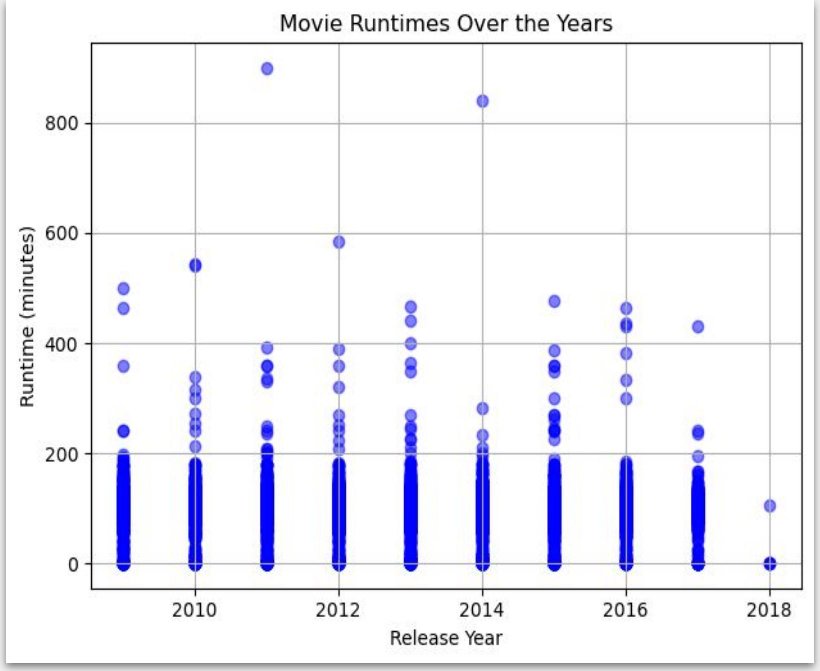
Question 3:
What is the trend of movie runtimes over the years?

```
#convert release date to datetime format
cleaned_movies_df["release_date"]=pd.to_datetime(cleaned_movies_df["release_date"], errors="coerce")
cleaned_movies_df["release_date"]=cleaned_movies_df["release_date"].dt.year
```

To answer the question, we converted the release date column to a year-only format.

Using movie_metadata, we created a DataFrame with release dates and runtimes for all movies.

Conclusion: Movie runtimes are slightly shorter but not significantly enough to raise concern.

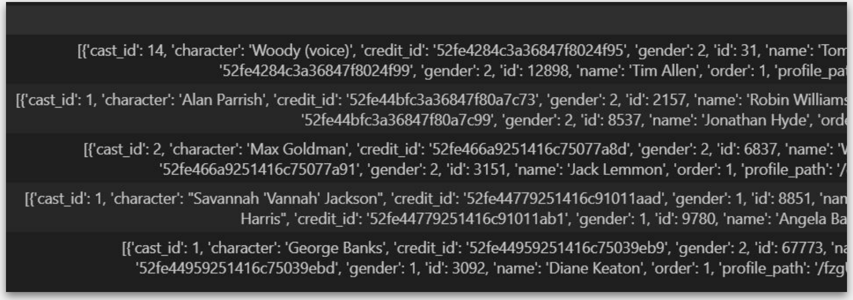


Question 4:
Which actors appear most frequently in high-rated movies?

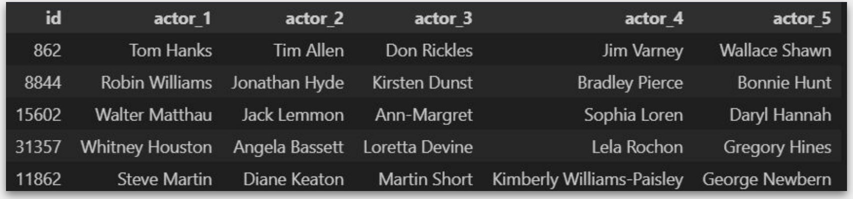
We worked with two datasets:
metadata_movies.csv and credits.csv.

To connect IMDb rating data with the actors listed in the credits dataset, we had to handle the cast column, which contained a list of dictionaries for each film.

Due to the large number of actors, we narrowed our focus to the top 5 actors for each movie and stored them in a separate column.



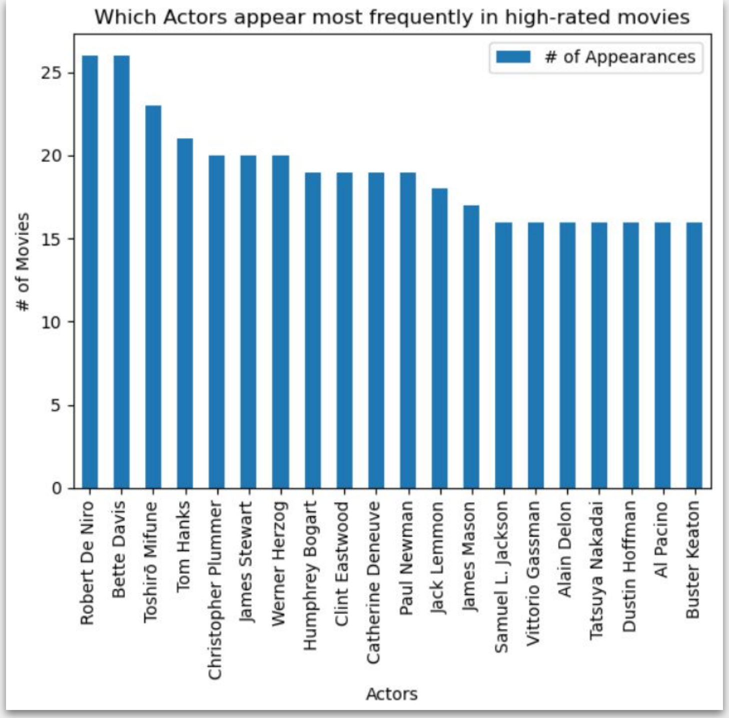
```
[{'cast_id': 14, 'character': 'Woody (voice)', 'credit_id': '52fe4284c3a36847f8024f95', 'gender': 2, 'id': 31, 'name': 'Tom
'52fe4284c3a36847f8024f99', 'gender': 2, 'id': 12898, 'name': 'Tim Allen', 'order': 1, 'profile_pat
[{'cast_id': 1, 'character': 'Alan Parrish', 'credit_id': '52fe44bfc3a36847f80a7c73', 'gender': 2, 'id': 2157, 'name': 'Robin Williams
'52fe44bfc3a36847f80a7c99', 'gender': 2, 'id': 8537, 'name': 'Jonathan Hyde', 'ord
[{'cast_id': 2, 'character': 'Max Goldman', 'credit_id': '52fe466a9251416c75077a8d', 'gender': 2, 'id': 6837, 'name': 'W
'52fe466a9251416c75077a91', 'gender': 2, 'id': 3151, 'name': 'Jack Lemmon', 'order': 1, 'profile_path': '/
[{'cast_id': 1, 'character': "Savannah 'Vannah' Jackson", 'credit_id': '52fe44779251416c91011aad', 'gender': 1, 'id': 8851, 'nan
Harris", 'credit_id': '52fe44779251416c91011ab1', 'gender': 1, 'id': 9780, 'name': 'Angela Ba
[{'cast_id': 1, 'character': 'George Banks', 'credit_id': '52fe44959251416c75039eb9', 'gender': 2, 'id': 67773, 'na
'52fe44959251416c75039ebd', 'gender': 1, 'id': 3092, 'name': 'Diane Keaton', 'order': 1, 'profile_path': '/fzg
```



id	actor_1	actor_2	actor_3	actor_4	actor_5
862	Tom Hanks	Tim Allen	Don Rickles	Jim Varney	Wallace Shawn
8844	Robin Williams	Jonathan Hyde	Kirsten Dunst	Bradley Pierce	Bonnie Hunt
15602	Walter Matthau	Jack Lemmon	Ann-Margret	Sophia Loren	Daryl Hannah
31357	Whitney Houston	Angela Bassett	Loretta Devine	Lela Rochon	Gregory Hines
11862	Steve Martin	Diane Keaton	Martin Short	Kimberly Williams-Paisley	George Newbern

Question 4:
Which actors appear most frequently in high-rated movies? (cont.)

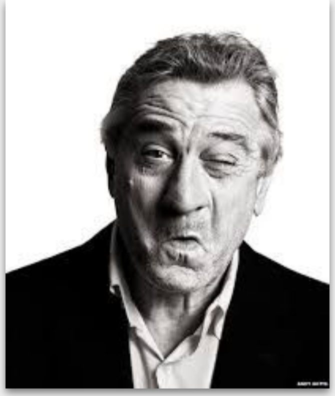
```
# Putting the 5 actors into their own column with a repeating ID
merged_movie_df = merged_movie_df.melt(
    id_vars=['id', 'imdb_id', 'original_title', 'release_date', 'title',
'vote_average', 'vote_count'],
    value_vars=['actor_1', 'actor_2', 'actor_3', 'actor_4', 'actor_5'],
    value_name='actor'
)

# Applying Vote Average Filter
filtered_merged_movie_df =
merged_movie_df[merged_movie_df['vote_average'] > 7]
actor_appearance_counts =
filtered_merged_movie_df['actor'].value_counts()

# Display the top 20 actors
actor_appearance_counts_df = actor_appearance_counts.reset_index()
actor_appearance_counts_df.columns = ['actor', 'appearance_count']
actor_appearance_counts_df.rename(columns={'actor': 'Actor',
'appearance_count': '# of Appearances'}, inplace=True)
actor_appearance_counts_df.head(20)
```



Question 4:
Which actors appear most frequently in high-rated movies? (cont.)

Robert De Niro starred in the movie, The Godfather Part II.

It was released in 18 Dec 1974 and it grossed a box office of $47,834,595 with an IMDb rating of 8.3.



```
imdb_id = filtered_actor.loc[filtered_actor['title']
== 'The Godfather: Part II', 'imdb_id'].values[0]

movie_actors = data['Actors'].split(", ")
target_movie_actor = movie_actors[1]
movie_title = data['Title']
movie_actor = data['Actors'][1]
movie_boxoffice = data['BoxOffice']
imdb_rating =
filtered_actor.loc[filtered_actor['title'] == 'The
Godfather: Part II', 'vote_average'].values[0]
release_date = data['Released']
```

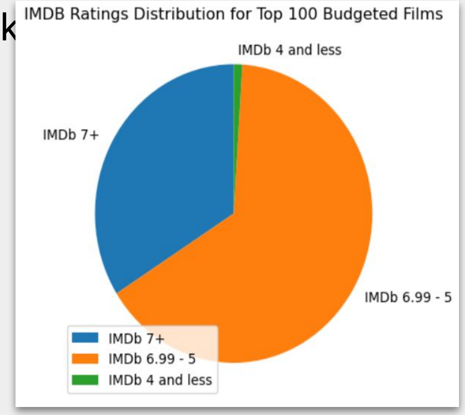
Question 5:
How does budget correlate with IMDb rating?

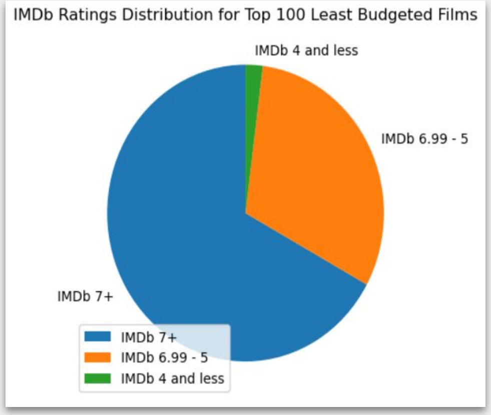
We worked solely in the `metadata_movies.csv`.

To understand trends in budget, we decided to work with the top 100 movies with the most and least budget.

The scores were grouped into 3 categories to showcase overall spread of the data.

Based on our finding, we were able to see that budgets with the least budget had a higher percentage of a higher rating.





Question 5:
How does budget correlate with IMDb rating? (cont.)

We used the OMDb API to gather additional information on our two movies.

Most budgeted film with a poor rating

The movie Independence Day: Resurgence had a budget of $165,000,000. Their Box Office was $103,144,286 with an IMDb rating of 4.9.

Plot: Two decades after the first Independence Day invasion, Earth is faced with a new extra-Solar threat. But will mankind's new space defenses be enough?

Least budgeted film with the okay rating

The movie Paranormal Activity had a budget of $15,000. Their Box Office was $107,918,810 with an IMDb rating of 5.9.

Plot: After moving into a suburban home, a couple becomes increasingly disturbed by a nightly demonic presence.

Question 5:
How does budget correlate with IMDb rating? (cont.)

We used the OMDb API to gather additional information on our two movies.

Most budgeted film with a poor rating

```
movie_title = data['Title']
movie_budget = topscores_4andless_df['budget'].iloc[0]
movie_boxoffice = data['BoxOffice']
imdb_rating = topscores_4andless_df['vote_average'].iloc[0]
movie_plot = data['Plot']

print(f'The movie {movie_title} had a budget of
${movie_budget:,}. Their Box Office was {movie_boxoffice}
with an IMDb rating of {imdb_rating}.')
print('')
print(f'Plot: {movie_plot}')
```

Least budgeted film with the okay rating

```
movie_title = data['Title']
movie_budget = bottomscores_6thru5_df['budget'].iloc[0]
movie_boxoffice = data['BoxOffice']
imdb_rating = bottomscores_6thru5_df['vote_average'].iloc[0]
movie_plot = data['Plot']

print(f'The movie {movie_title} had a budget of
${movie_budget:,}. Their Box Office was {movie_boxoffice}
with an IMDb rating of {imdb_rating}.')
print('')
print(f'Plot: {movie_plot}')
```

Question 6
Which countries produce the most amount of movies in each genre?

The focus of the last 3 questions was on movie production as well as financial performance. We decided to focus on the data set metamovies_dataset.csv as it contained information over a vast arrays of categories.

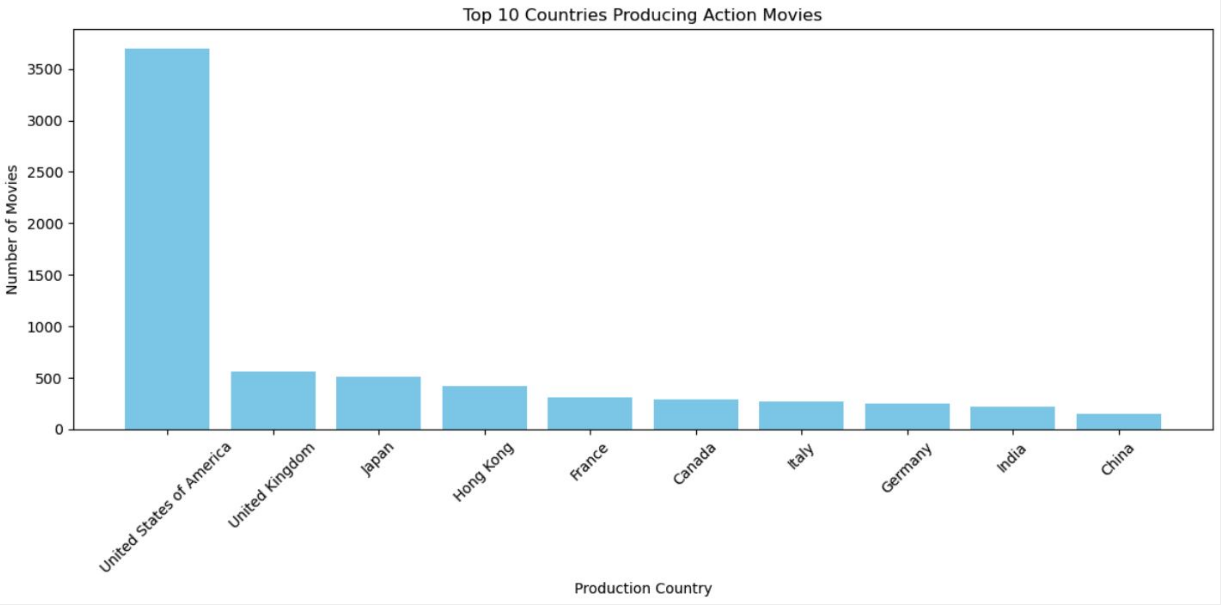
Cleaning this data allowed us to narrow our focus on the most relevant columns and gain insights into all of these questions.

```
# Drop columns with unneeded data.

clean_movies_metadata_df = movies_metadata_df.drop(columns=["adult",
"poster_path", "production_companies", "belongs_to_collection",
"homepage", "original_language", "original_title", "overview",
"release_date","runtime", "spoken_languages", "status",
"tagline","video"])
clean_movies_metadata_df
#drop rows with missing values

clean_movies_metadata_df=clean_movies_metadata_df.dropna()
clean_movies_metadata_df=clean_movies_metadata_df.dropna(axis=1,)

clean_movies_metadata_df
```

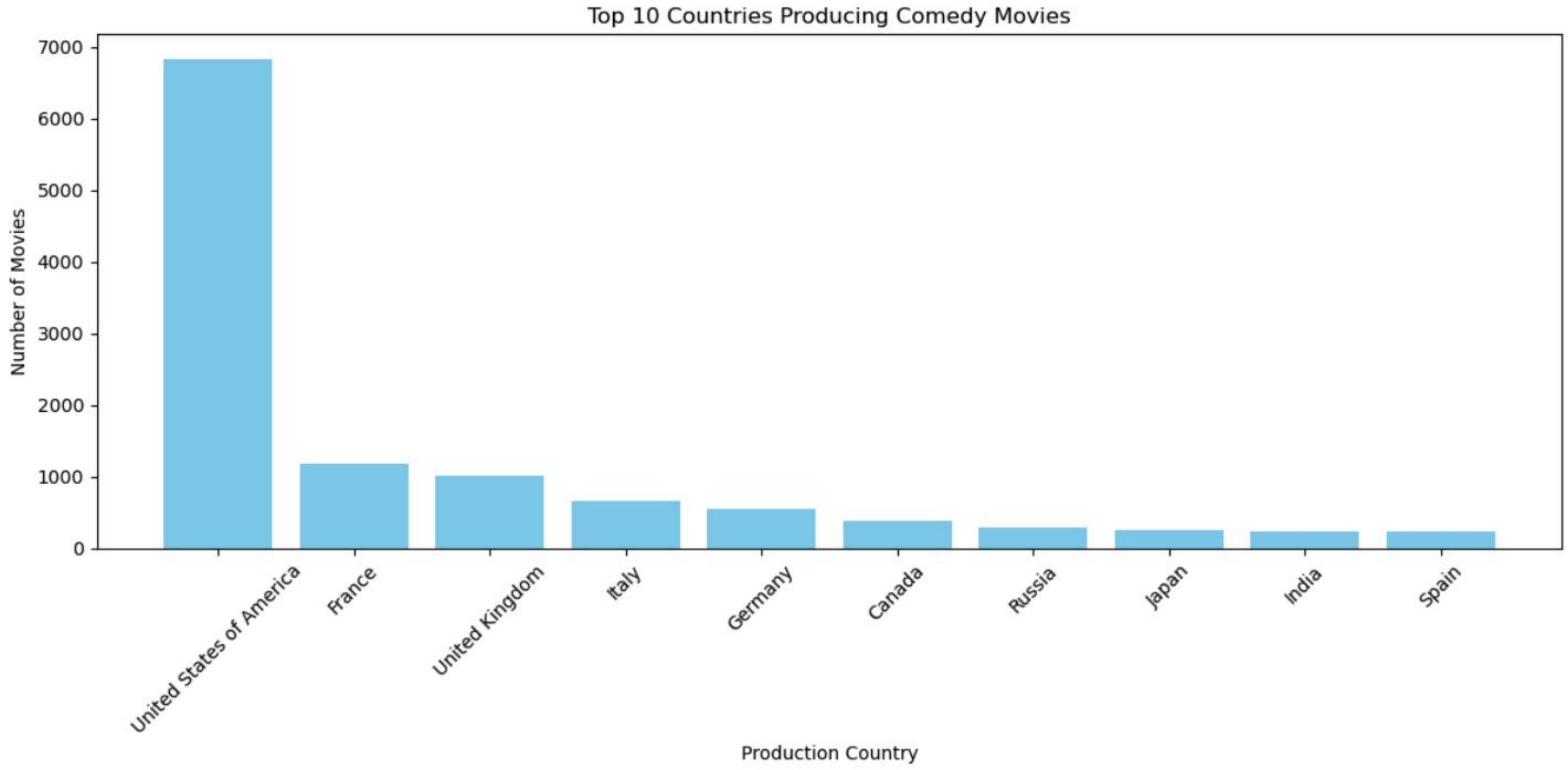
Result

Action Movies

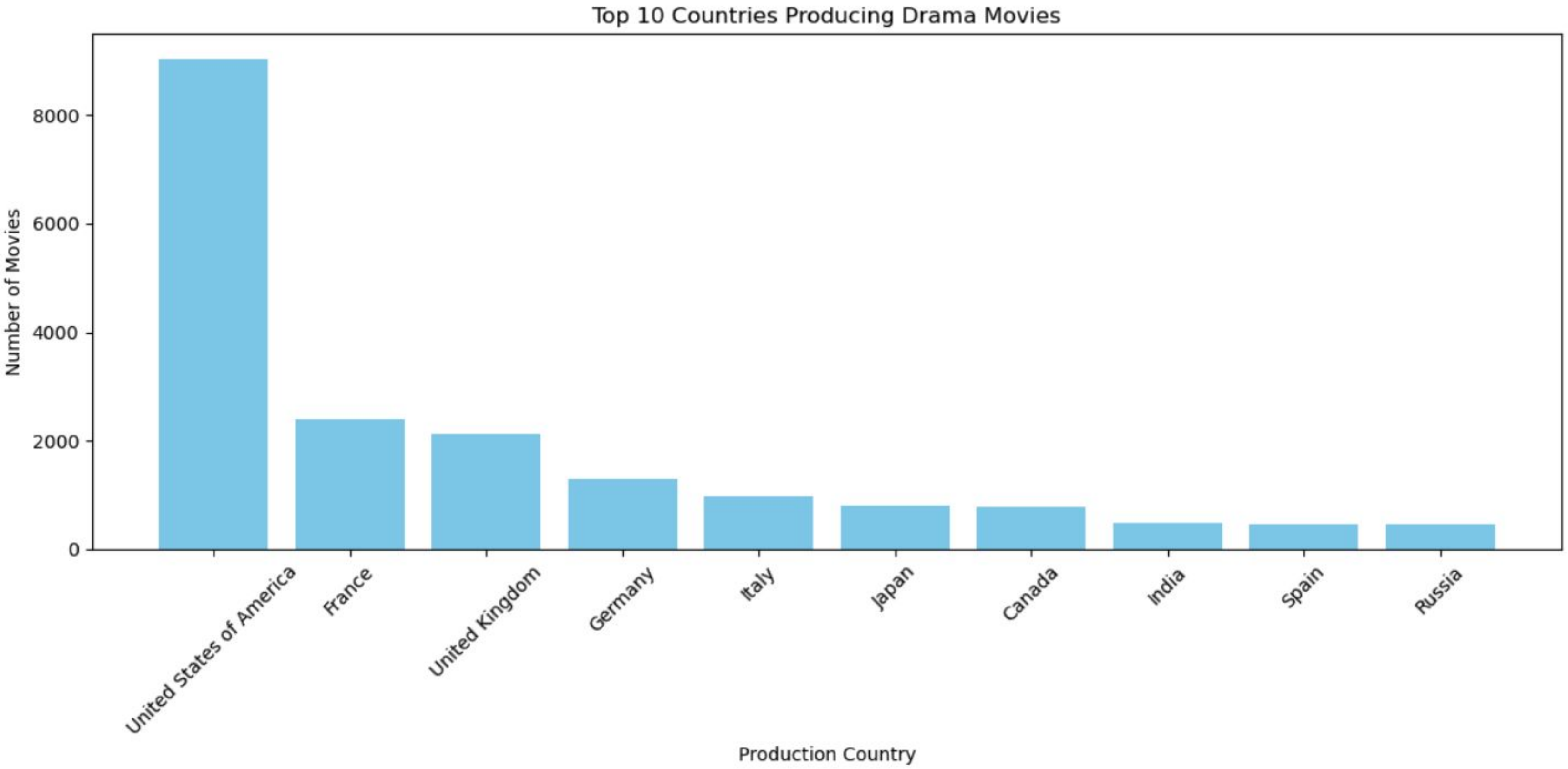
The U.S. leads in the amount of movie productions across every category but the rest of the genres are mainly produced in France, Italy, Germany, Japan and The U.K.



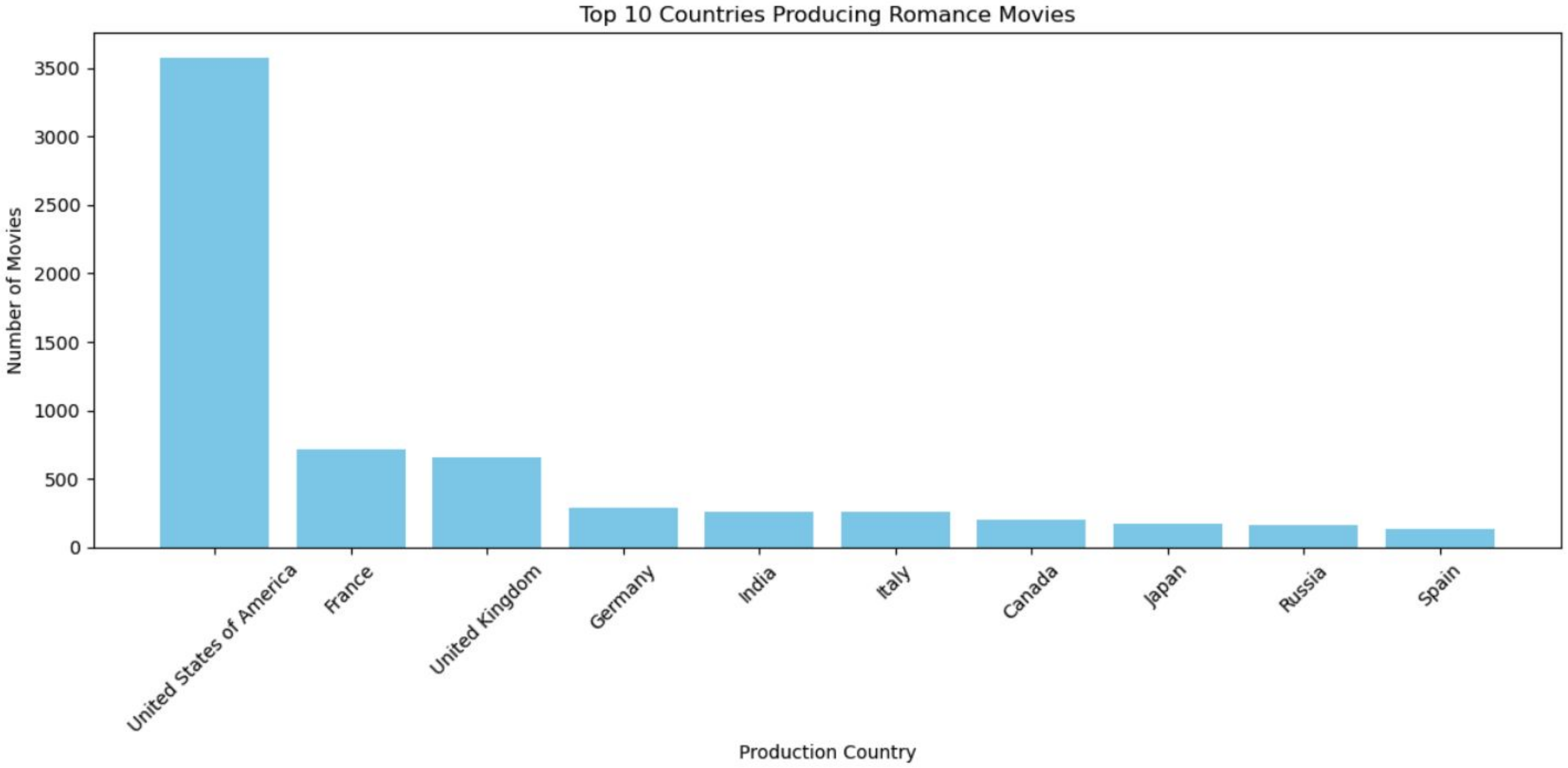
Comedy


Top 10 Countries Producing Comedy Movies
Number of Movies
7000
6000
5000
4000
3000
2000
1000
0
United States of America
France
United Kingdom
Italy
Germany
Canada
Russia
Japan
India
Spain
Production Country

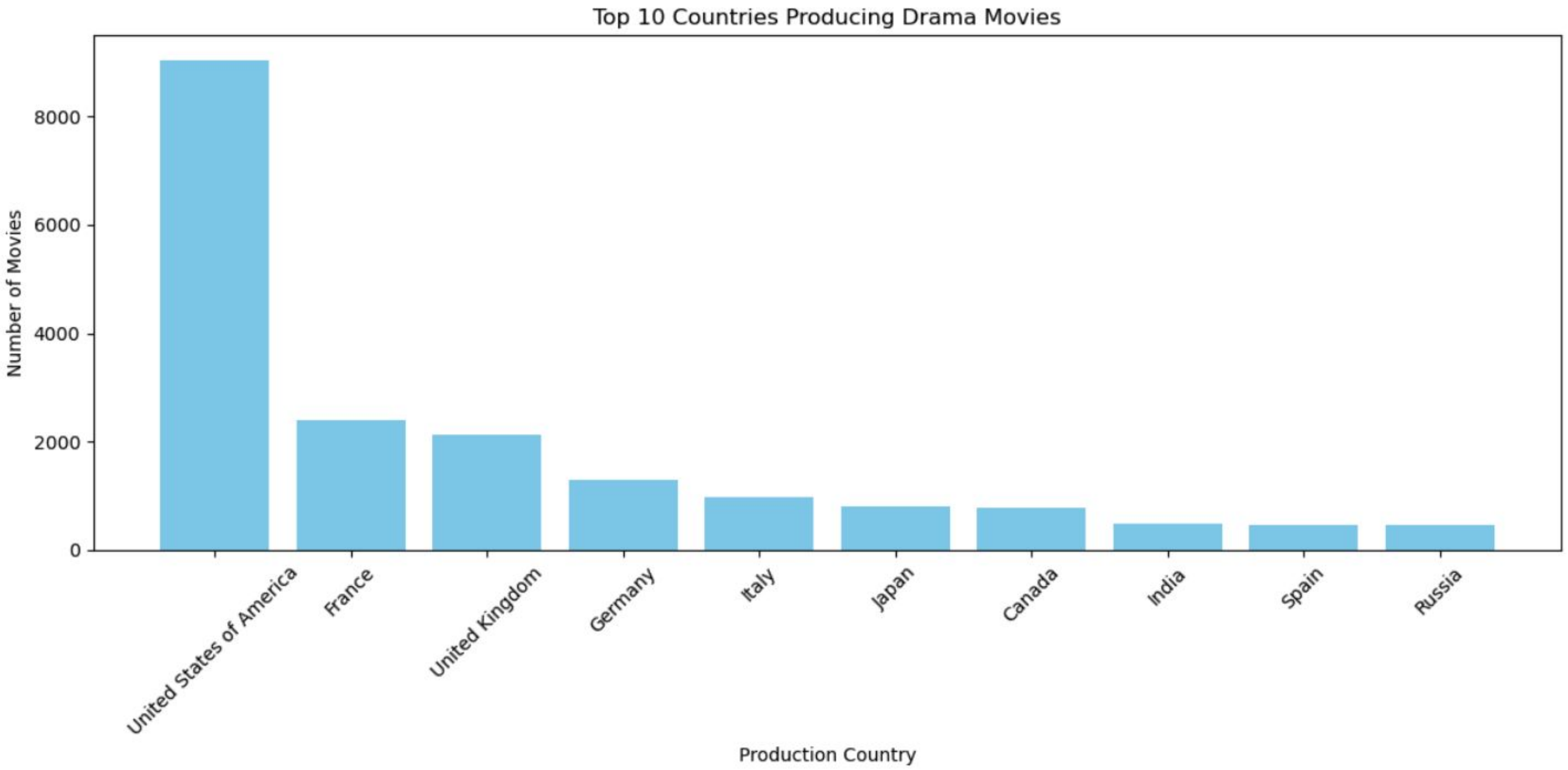
Drama


Top 10 Countries Producing Drama Movies
Number of Movies
0
2000
4000
6000
8000
United States of America
France
United Kingdom
Germany
Italy
Japan
Canada
India
Spain
Russia
Production Country

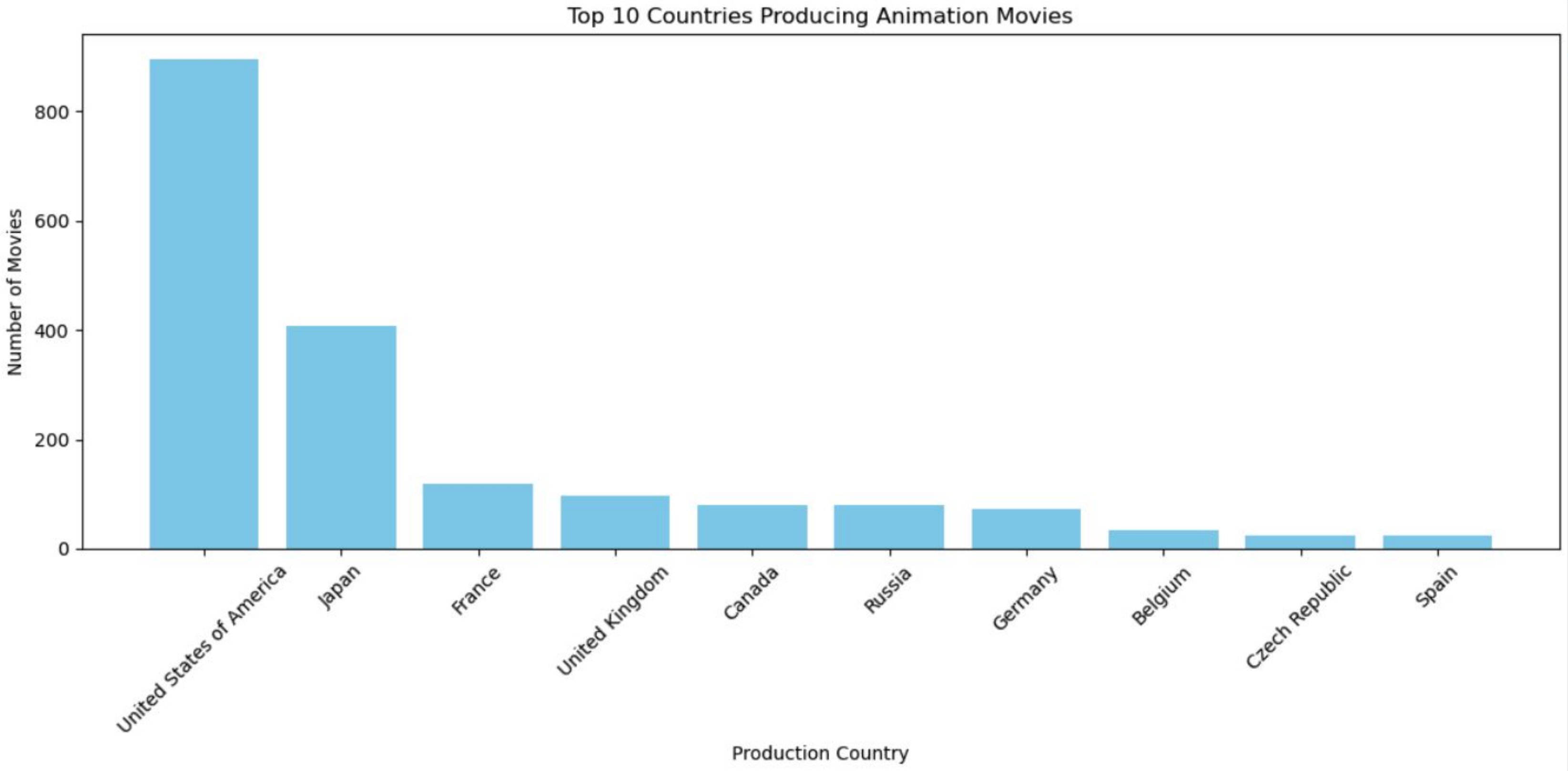
Romance


Top 10 Countries Producing Romance Movies
Number of Movies
0
500
1000
1500
2000
2500
3000
3500
United States of America
France
United Kingdom
Germany
India
Italy
Canada
Japan
Russia
Spain
Production Country

Drama


Top 10 Countries Producing Drama Movies
Number of Movies
0
2000
4000
6000
8000
United States of America
France
United Kingdom
Germany
Italy
Japan
Canada
India
Spain
Russia
Production Country

Animation


Top 10 Countries Producing Animation Movies
Number of Movies
0
200
400
600
800
United States of America
Japan
France
United Kingdom
Canada
Russia
Germany
Belgium
Czech Republic
Spain
Production Country

Question 7

On average, which movie genre generates the most revenue?

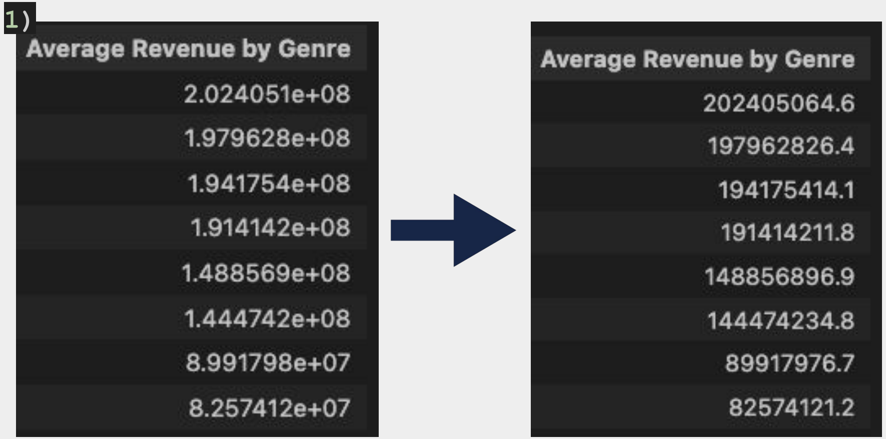
With the increased sophistication and uses of new technologies, movies today look and feel better than ever. We wanted to know how this is reflected in the revenue generated by each genre.

```
cleaned_metadata_df =
cleaned_metadata_df[(cleaned_metadata_df
!=0).all(axis=1)]
```

```
revenue_by_genre_df = cleaned_metadata_df
.groupby("Genres")["Revenue"].mean().reset_index()
```

Working with the same data set metamovies_dataset.csv we did some more cleaning, grouped movies by genres and calculated the mean of the different categories.

```
revenue_by_genre_df["Average Revenue by Genre"] =
round(revenue_by_genre_df["Average Revenue by Genre"],
1)
```

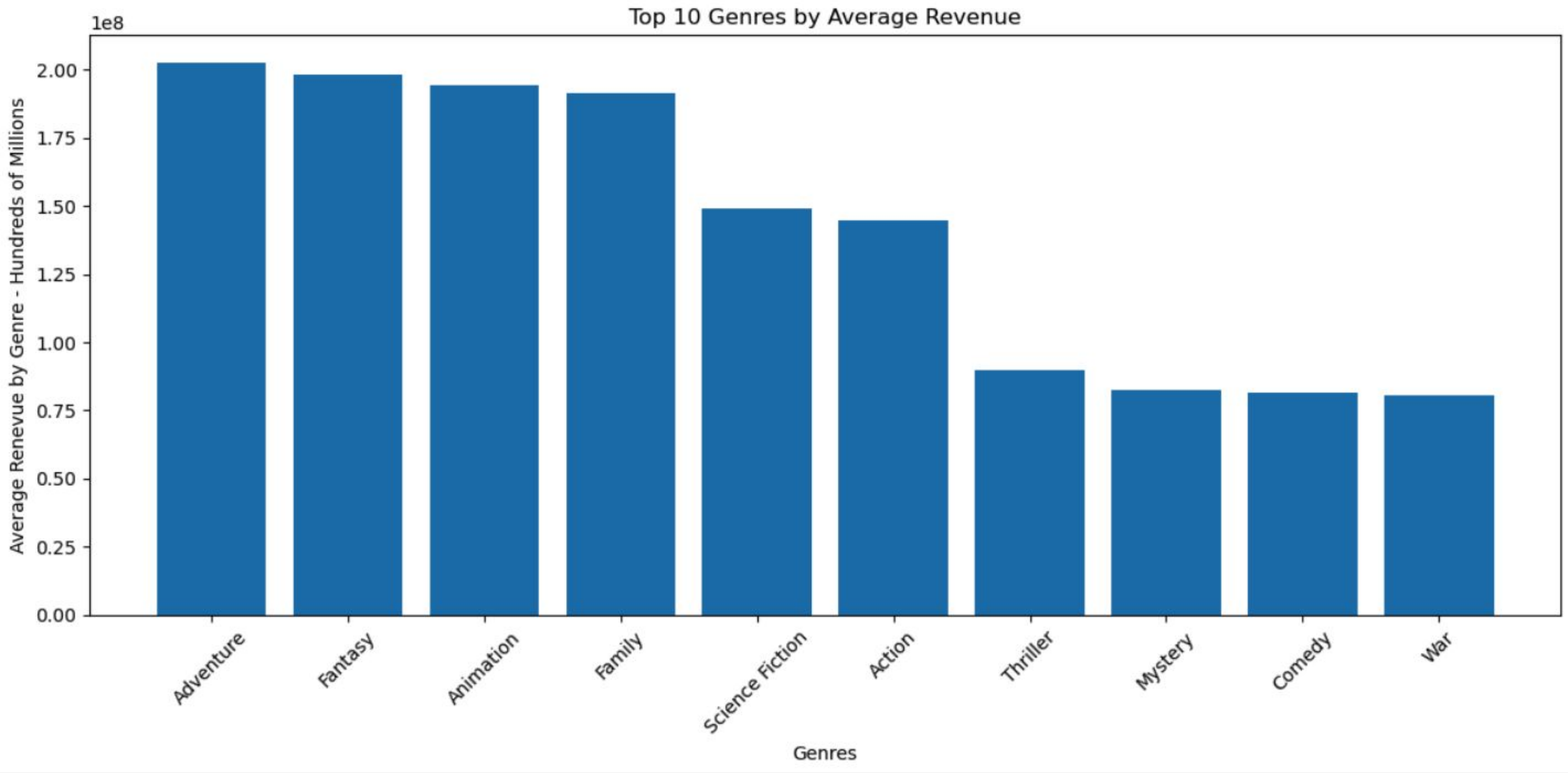


```
Average Revenue by Genre
            2.024051e+08
            1.979628e+08
            1.941754e+08
            1.914142e+08
            1.488569e+08
            1.444742e+08
            8.991798e+07
            8.257412e+07
```

```
Average Revenue by Genre
             202405064.6
             197962826.4
             194175414.1
             191414211.8
             148856896.9
             144474234.8
              89917976.7
              82574121.2
```

Result

While comedy brings in as much revenue as mystery films, the adventure genre has doubled their performance in recent decades.

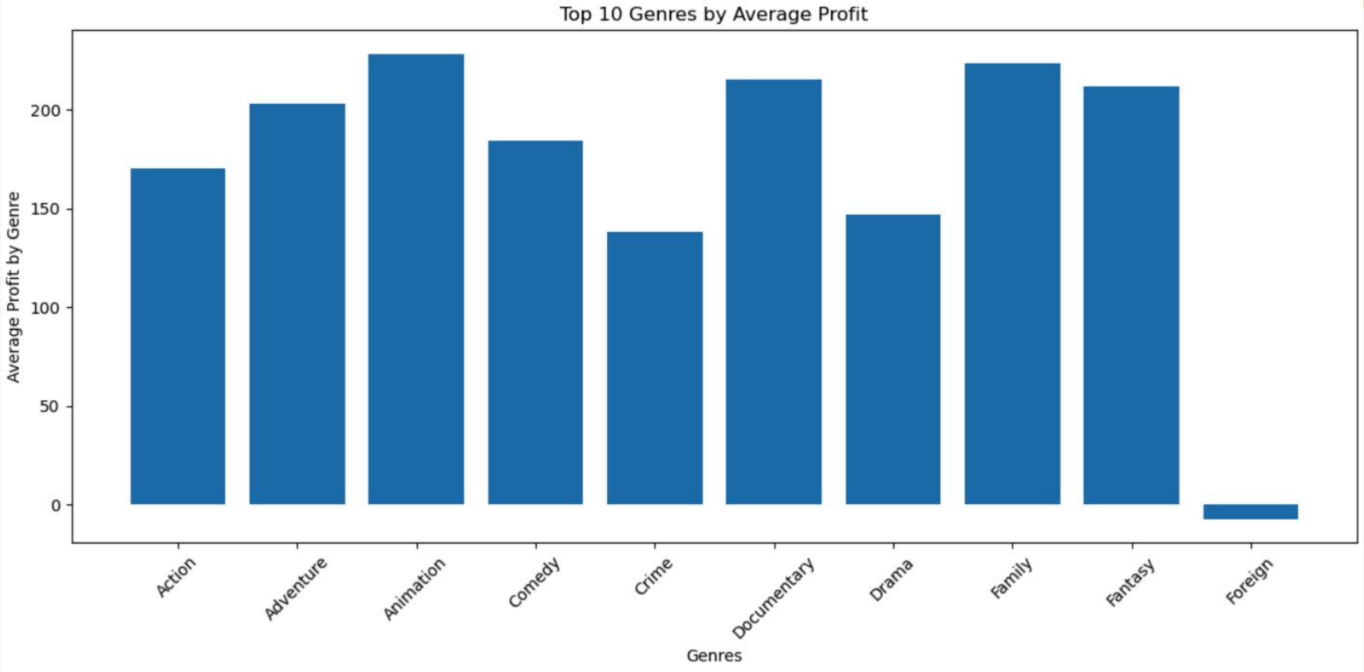


Question 8

Does a genre's revenue correlate to its profitability?

Result: They do!

While movie production is highly profitable across the board, the family, adventure and animation genres tend to bring the most revenue as well as the highest average profits at over 200%.



fin.