**Data Processing:**

### Dataset Summary:

* The dataset was imported from a csv file, containing 57 636 values, with 12 columns and 4803 records.The columns include:

| **Feature** | **Data Type** | **Description** |
| --- | --- | --- |
| budget | Integer | The cost of production of the movie. |
| genres | List | The type of movie. |
| homepage | String | Web address of the movie. |
| id | Integer | Unique identifier for each movie feature in the dataset. |
| keywords | List | Short description of movie. |
| original\_language | String | Language the movie was shot in. |
| original\_title | String | The name of the movie |
| overview | String | Description of the movie |
| popularity | Float | First week viewings of the movie. |
| production \_companies | List | Companies that were involved in the production of the movie |
| production\_countries | List | Countries that the movie was shot in |
| release\_date | String | Release date of the movie |
| revenue | Integer | The amount of money the movie made. |
| runtime | Float | The duration of the movie |
| spoken\_languages | List | Languages spoken in the movie |
| status | String | Whether the movie has been released or not |
| tagline | String | Catch phrase for the movie |
| title | String | The name of the movie |
| vote\_average | Float | Viewer average rating out of 10 |
| vote\_count | Integer | The total number of people who have given the movie a rating out of ten |

### Data Cleaning:

1. **Duplicated Values:**
   1. Identify the total number of duplicated values by chaining the ***.duplicated()*** method and the ***.sum()*** methods to return the total number of blank values within the dataset.
   2. There are no duplicate values in the dataset.
2. **Null Values:**
   1. Identified the total number of null values by applying the numpy function ***np.count\_zero()*** and passed the dataframe with the method ***.isnull()*** chained with the ***.sum()*** method into the function.
   2. There were no null values within the dataset.
3. **Outlier Values:**
   1. The outliers were visualized using the seaborn library’s function, ***sns.boxplot()***
   2. The user defined function, ***outlier\_imputation().*** This function implemented the IQR to find the total number of outliers in each numeric column. This would then be outputted to find
   3. The same function was then responsible for imputing the outlier values, by either the mean, median, or upper and lower limits that were defined within the function.
   4. The distribution of the features were then visualized using the ***sns.histplot()*** to compare the distribution of the feature before the imputation, and after the imputation

### Data Transformation:

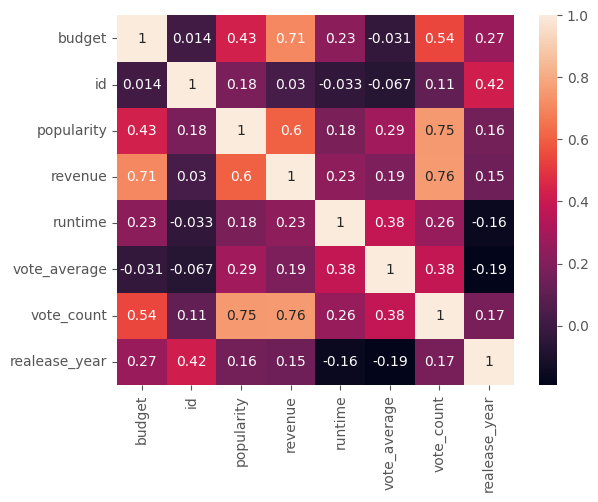
1. **Data Normalization:**
   1. Applied Min-Max normalization to scale numerical features to the range [0, 1].

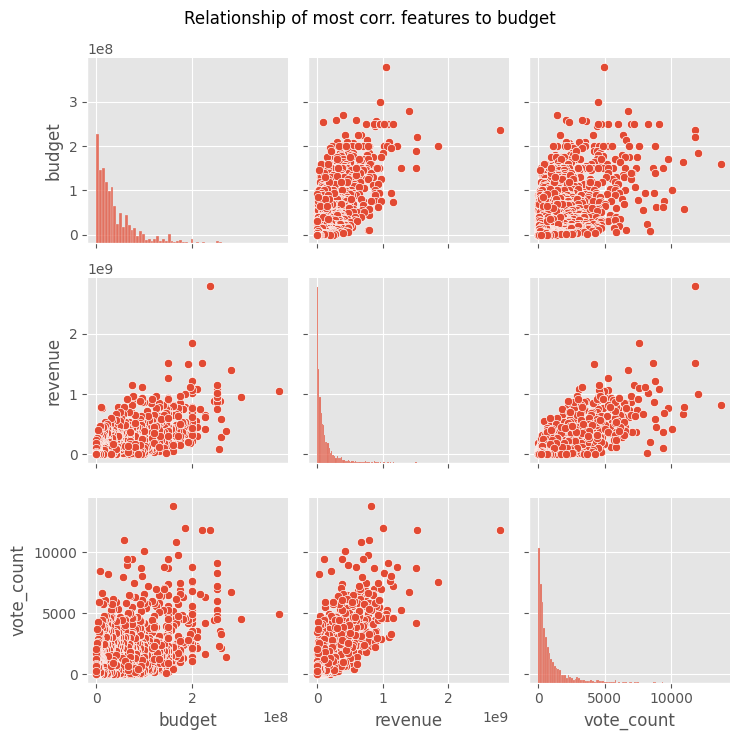
### Data Engineering:

1. **Feature Engineering:**
   1. Read in geospatial data, giving information about the geographical shape of a region, in this case, the states of America
   2. The original dataset did not have a feature that we could link to the geospatial dataset.
   3. Read in the dataset scraped from wikipedia about the state, region and state code, which matches the codes of the states in the original dataset stored as df.
   4. Filter the webscraped dataset from wikipedia stored as state\_code\_df to only include the state column and the state code, and store as dictionary, in the variable state\_code\_dct.
   5. Store a copy of the original dataframe using the ***.copy()*** method. As df3.
   6. Add a new feature, Code, into df3, based on the column ‘City’. Apply ***.map()*** and pass state\_code\_dct as argument so the value from the dictionary is returned based on the value in the ‘City’ column matching the key of the dictionary

### Data Stories and Visualizations:

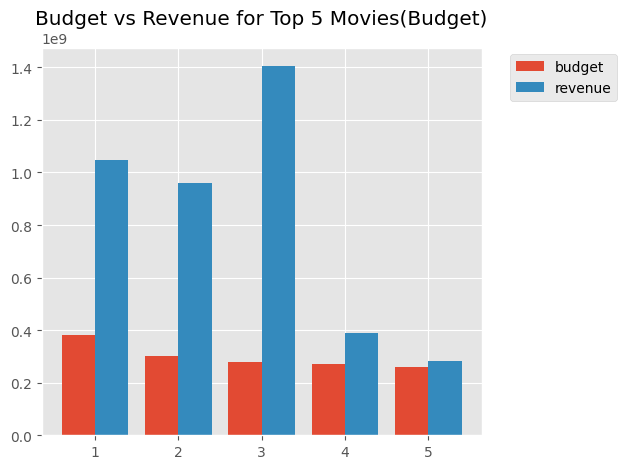
**Relationships and Distributions:**





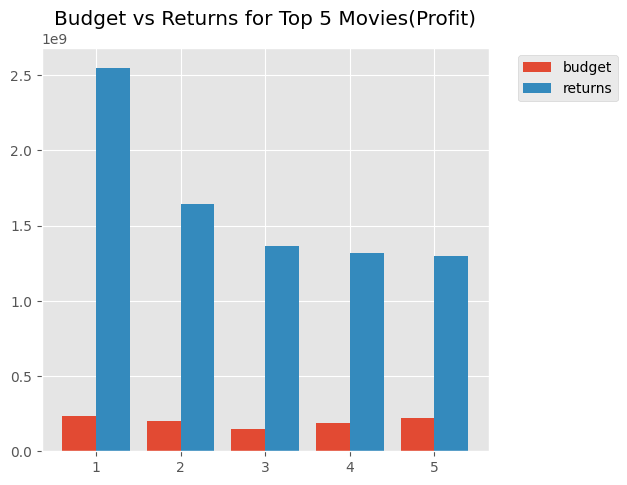
* Based on the correlation matrix, we can see that the features most correlated to budget, other than itself, are revenue and vote\_count. We then use the pairplot to view the relationship between these features with regards to the distribution of the feature, and the strength of the correlation. This is seen with the distribution of the points around the gradient line of the scatter plot representing the relationship between the features.

**Budget of the top 5 revenuing movies:**

****

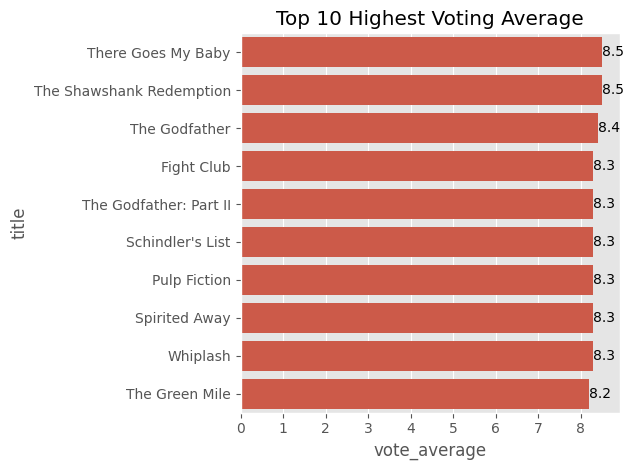
* The movies on the x-axis: 1. Pirates of the Caribbean: On Stranger Tides, 2. Pirates of the Caribbean: At World's End, 3. Avengers: Age of Ultron, 4. Superman Returns, 5. John Carter
* The movies with the highest revenue is ***Avengers: Age of Ultron***

**The top 5 most profitable movies:**

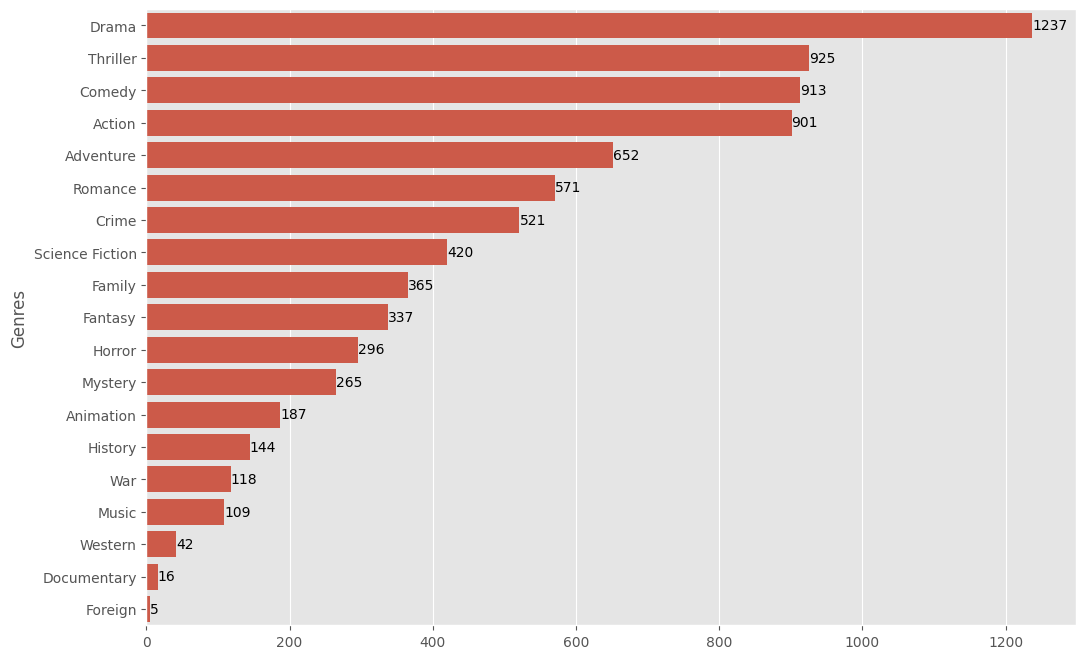


* The movies on the x-axis: 1. Avatar, 2. Titanic, 3. Jurassic World, 4. Furious 7, 5. The Avengers
* The most profitable movie from the dataset is ***Avatar***

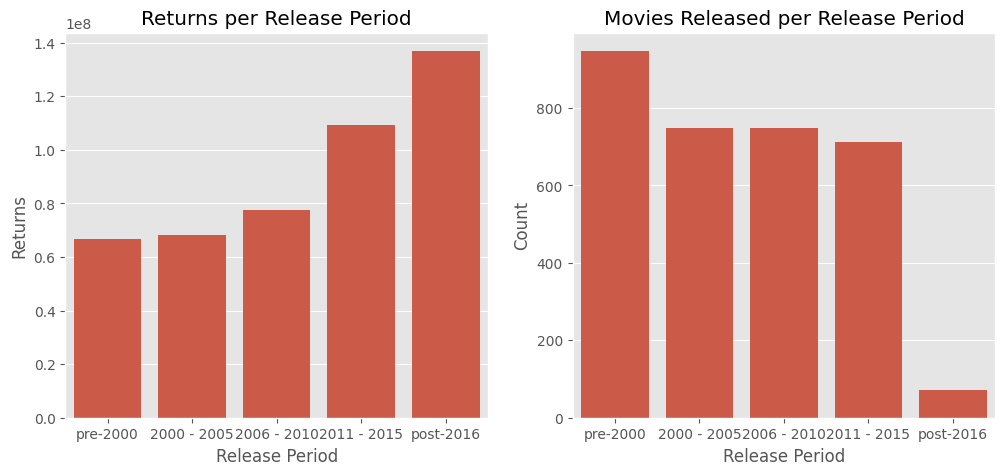
**The top 5 most profitable movies:**

* Based on the barplot, we can see that the highest rated movies on average, in ascending order: 1. There Goes My Baby, 2. The Shawshank Redemption, 3. The Godfather, 4. Fight Club, 5. The Godfather: Part II, 6. Schindler’s List, 7. Pulp Fiction, 8. Spirited Away, 9. Whiplash, 10. The Green Mile

**The popularity of genres:**

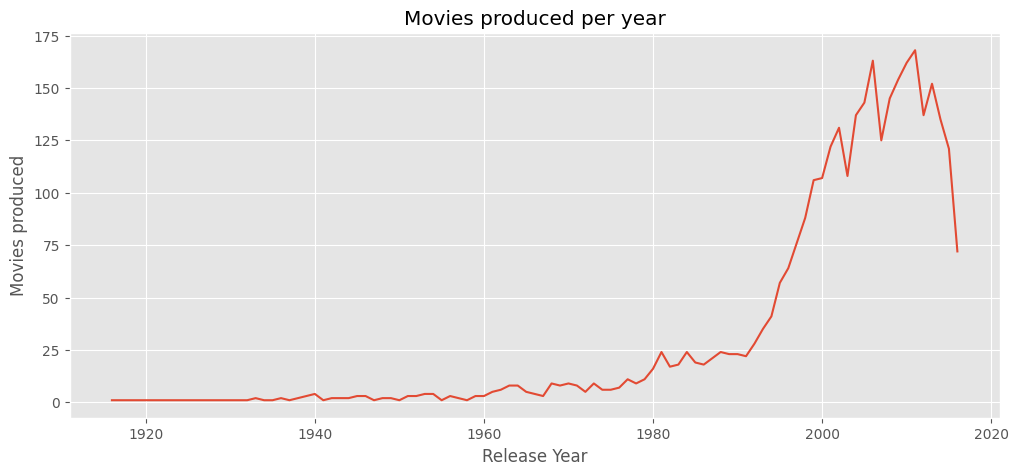


* From the barplot, we can see that most popular genre is Drama

**The Revenue by release period and Total movies released per period:**

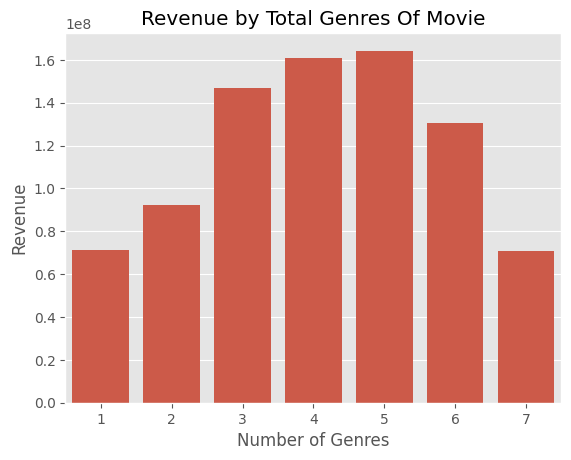
* From the plot, we can surmise that the total returns from movies have steadily increased over each period of time. The interesting thing is that even though the total number of movies released during the 'post-2016' period shows to be the lowest with regards to releases, it is during this period where we find the most profitable movies coming out.
* Interestingly as well, from the above output, we can also see that from the '2006 - 2010' period, there has been a steady increase in the ratio of action and adventure movies coming to the fore. According to Chatgbt, Science Fiction, Fantasy, Superhero, Action/Adventure, Animated Films and Horror, are the genres which leverage CGI (Computer-Generated Imagery) (Which movie genres uses the most cgi on average?, 2024). Based on the top 5 most profitable movies produced in cell 37 and 38, we can see that there is some truth in this.
* Adventure, action, sci-fi and fantasy movies take the longest to reach the theater from the first announcement, averaging at least 1000 days. These total days include first announcement, pre-production, shooting, and post-production. These categories also experience the longest first announcements, this could be because of many reasons, but one of the major reasons is to signal their intent to rival studios. This may help them secure a prime release slot in the calendar or ensure there is no competition for their movie’s topic or theme. (Stephen Follows, 2018)

**Total number of movies produced per year:**



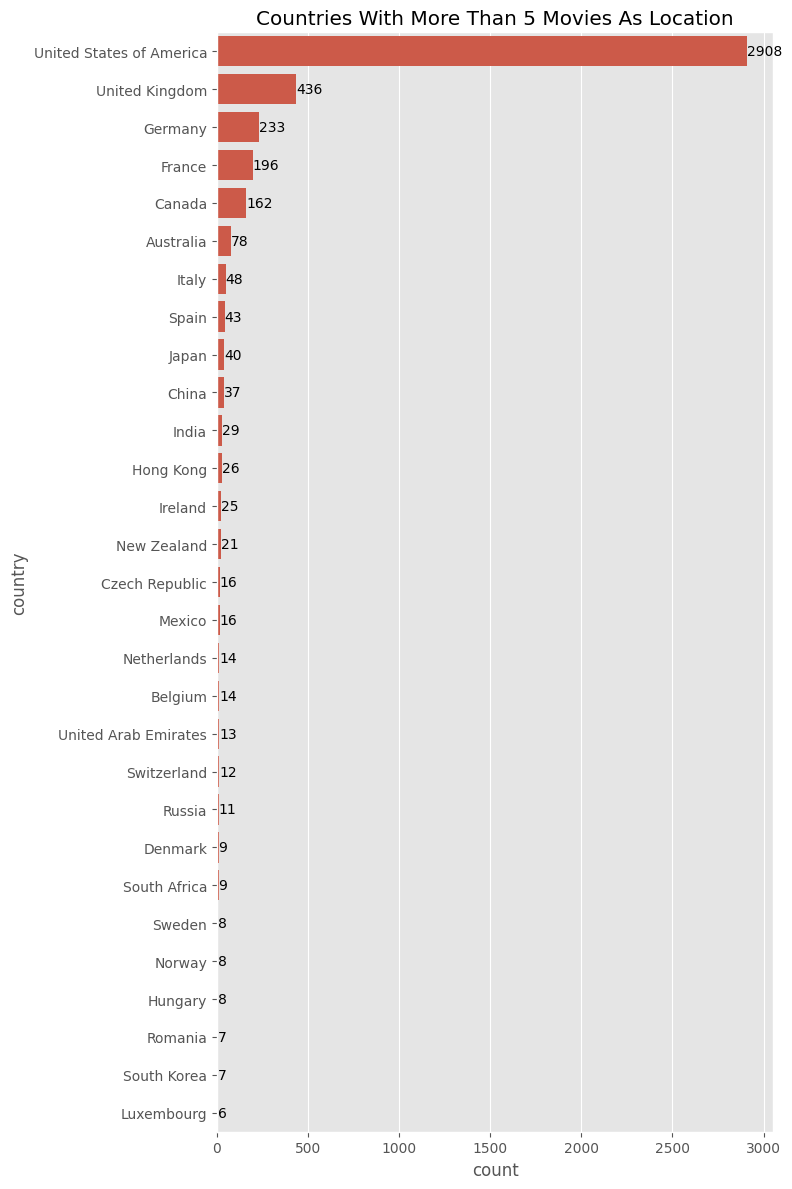
* We can see that there has been an increase gradually over time with regards to the total number of movies released per year. The fall off at 2016 is due to the dataset not including movies that were released at the end of the year.

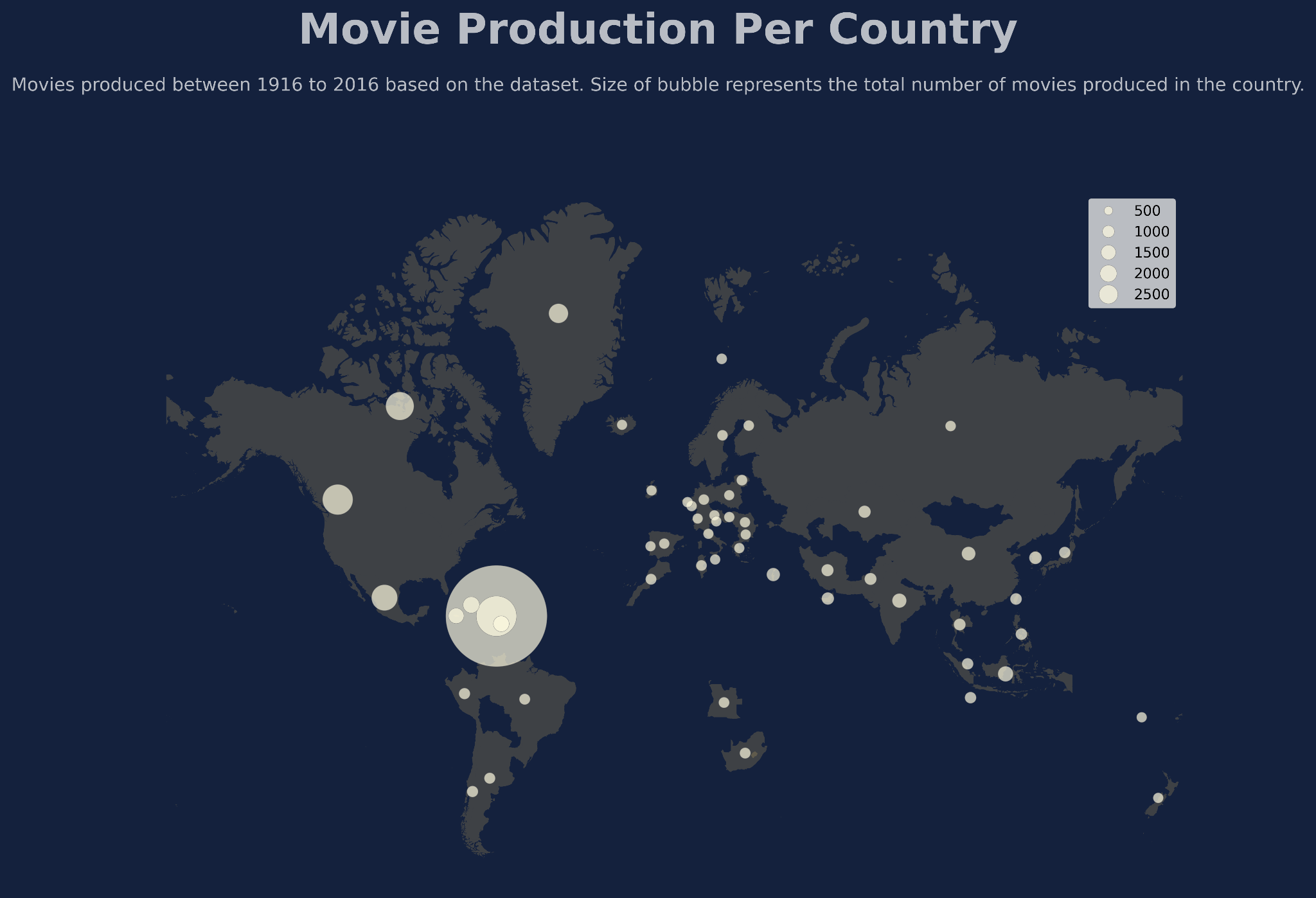
**The average revenue returned based on the genres a movie is characterized in:**



* From the bar plot above, there seems to be a threshold where the total number of genres suggests that movies that include multiple genres, tend to get to a larger audience, and that it could result in greater revenue. However, this relationship could also be further researched to find if there is a combination of movies that return a greater revenue than others.

**Countries With More Than 5 Movies As Location:**





* From the barplot we can see that from the dataset, the most popular country for mobile features, is the United States. In an article by Kyle Deguzman (2024) that in Hollywood, the term 'big studio' refers to a film production company that not only produces films but also distributes them. These movie studios have large, dedicated facilities, significant resources, and a substantial workforce. They have the capacity to handle all aspects of film production, from scriptwriting and filming to post-production and distribution.
* With the description of a big studio in mind, based on another article written by Ria Puneyani (2023), based on the total revenue generated from their movies at the time, the 10 biggest production companies in the world are, 1. \_`Universal Pictures`\_ with a total revenue of $11.62 billion, 2. \_`Warner Bros`\_ with a total revenue of $44.3 billion, 3. \_`The Sony Pictures Motion Picture Group`\_ with a total revenue of $7.16 billion, 4. \_`Walt Disney Pictures`\_ with $82.7 billion total revenue, 5. \_`Paramount`\_ with a total revenue of $30.15 billion, 6.\_`Lionsgate`\_ with a total revenue of $35.97 billion, 7. \_`20th Century Fox`\_ with a total of $47 million, 8. \_`Weinstein Company`\_ comes in 8th with a total revenue of $33.9 million, followed by \_`Metro-Goldwyn-Mayer Studios`\_ at 9th with a total revenue of $13.1 billion, and finally, at 10. \_`DreamWorks Studios`\_ with a total revenue of $915.9 million.
* Therefore, we can deduce that having the top ten movie production companies being based in the United States of America, it makes sense why the country is the most popular site for movie production.

**Bibliography:**

1. Which movie genres uses the most cgi on average? (2024), Chatgbt GPT-4o, 28 July, Link:

<https://chatgpt.com/share/a3b073c0-da04-4e87-bdce-d5dcdc9ccad3>

2. Stephen Follows (2018), 'How long does the average Hollywood movie take to make?', \_Stephen Follows Decoding The World Through Data\_, 7 may, Available at:

<https://stephenfollows.com/how-long-the-average-hollywood-movie-take-to-make/>

3. Deguzman, K. (2024), 'What are the Major Film Studios — Hollywood’s Big Five', \_Studiobinder\_, 27 January. Available at:

<https://www.studiobinder.com/blog/what-are-the-major-film-studios/#:~:text=The%20'Big%20Five'%2C%20consisting,industry%20over%20the%20last%20century.>

(Accessed: 10 August 2024)

4. Puneyani, R. (2023), 'Top 10 Movie Studio Companies in the World', \_StartupTalky\_, 08 September. Available at:

<https://startuptalky.com/top-movie-studio-companies/>

(Accessed: 10 August 2024