PRESENTER: JOSEPHINE GATHENYA

COURSE:DATA SCIENCE

DATE: 5/11/2023

PHASE ONE PROJECT

PROJECT STATEMENT

- Microsoft Company wants to join the creative movie industry by creating a new movie studio. They however do not have a background in creating movies and, would like to know, which films are performing the best at the Box Office.

- The objectives of this project are to use data analytic tools and techniques to gain useful insights and decision-making that will help Microsoft decide what type of films to create.
- Important information I will need to know about the best-performing films:
- The genres, the film titles, the ratings based on the viewership votes, the budgets used, and how much the films made as profit both locally and internationally.

FILES USED FOR ANALYSIS

- The following files were used:
- IMDb.title.basics
- IMDb.title.ratings
- BOM.movie_gross
- TN.movie_budgets

METHODOLOGY

- The necessary libraries were imported for analysis and visualization i.e. Pandas,numpy,matplotlib, and Seaborn.
- Data was retrieved from the 4 CSV files using the Pandas module, namely:
- IMDb.title.basics, IMDb.title.ratings, BOM.movie_gross and TN.movie_budgets
- Data cleaning was done before analysis by dealing with missing values which involved dropping rows where there were few missing values and replacing missing values with the median value where there were a lot of missing values.
- Comma values and dollar signs were also removed during the cleaning exercise.
- Data was converted from object values to float or integer values where calculations needed to be done e.g. mean, median, and sum of values in the data frames.
- Data filtering was done using join and group by statements.
- Data analysis was then done with a focus on the best-performing films based on their genres, their ratings based on viewers' votes, the budgets used, and the profits made.

FINDINGS

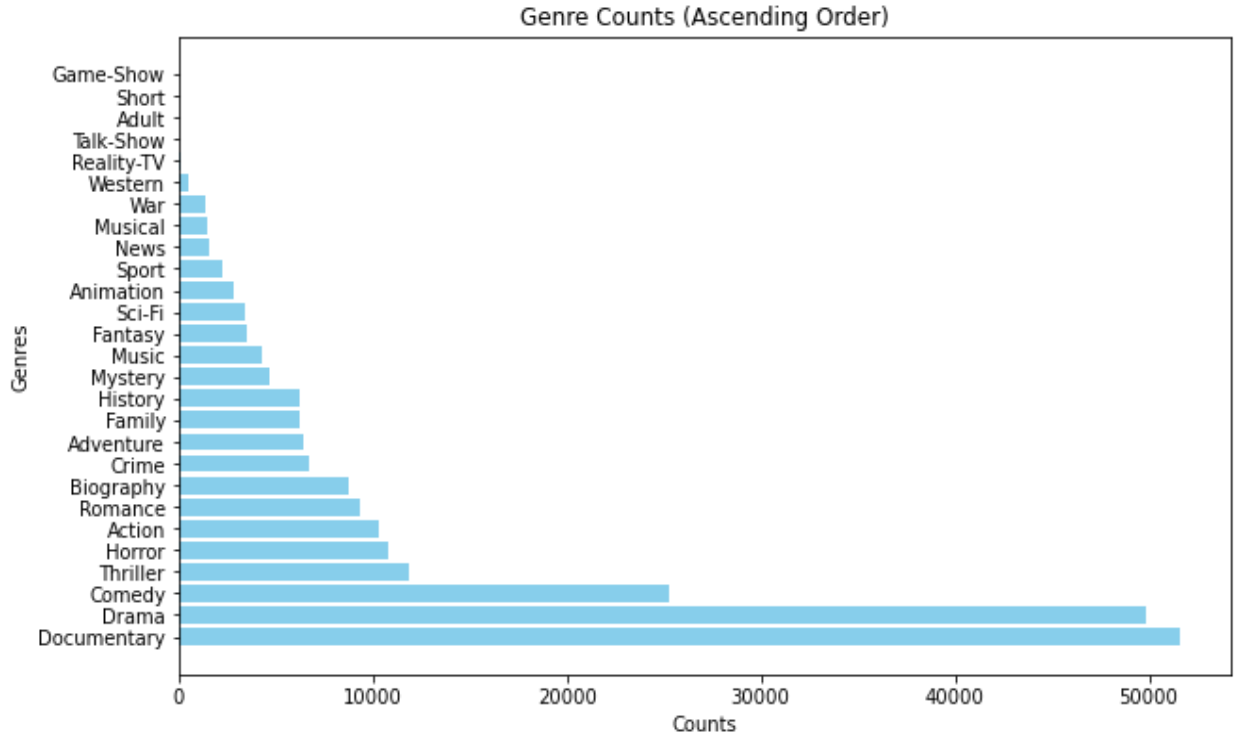
- The following were the findings :
- The Top 6 best-performing genres based on the frequency of occurrence from the 'title basics' dataset were as follows, in descending order:

GENRE	NUMBER OF OCCURENCES	PERCENTAGE OCCURENCE
Documentary	51640	36.7%
Drama	49883	35.4%
Comedy	25312	18.0%
Thriller	11883	8.4%
Horror	10805	7.7%
Action	10335	7.3%

- The total films analyzed were 140734 in this dataset.

...MORE FINDINGS

- The following bar chart shows the frequency of occurrence of all the films in the title basics dataset:

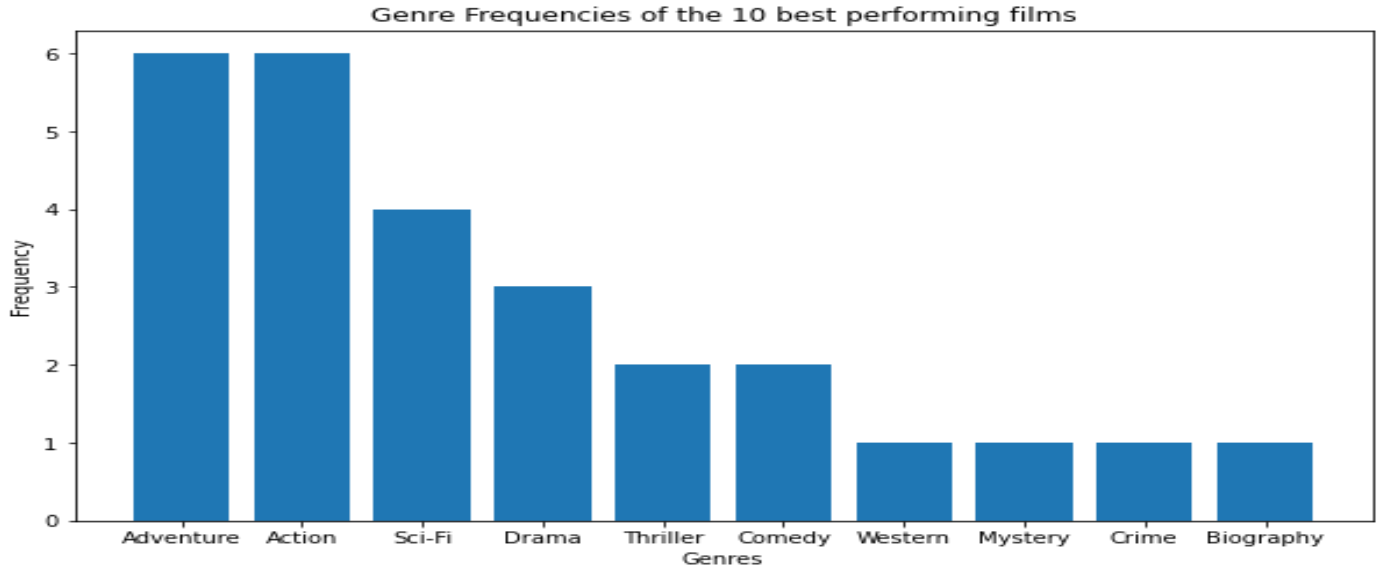


Film ratings findings

- The title basics dataset was then joined with the 'title ratings' dataset, using the 'tconst' column.
- An inner join was used to retrieve only the matching records from the two datasets.
- This was done to find out the highest-performing films based on the ratings and the number of viewers who cast their votes.
- It was noted that looking at the best films based on the ratings only would create a bias because some ratings that were labeled high ratings had fewer votes e.g. 5 viewers all giving a high rating.
- 73052 matching records were retrieved.

- **9326/73052 films had a high rating of 8.0 and above.**
- The ‘numvotes’ column in the merged data frame was then sorted out in descending order to show the films which had the highest votes cast.
- The top 10 films were retrieved and it was found that 9 of them had a high rating and 1 film had a medium rating of 7.2.
- Data showed that the 10 best performing films had higher viewership and were given a higher rating.
- The focus was now shifted to these films for further analysis.

- The genres of these 10 films were analyzed and the frequency of their occurrence was investigated.
- The findings showed that viewers most liked genres were Action, Adventure, and Sci-Fi films.
- The bar chart below shows the genre frequencies of the 10 best films:



Budgets and profits findings

- The ‘bom movie gross’(right df) dataset containing the foreign gross and domestic gross was joined to the ‘movie budgets’ dataset(left df) which contained the product budgets, domestic gross, and worldwide gross.
- The ‘title’ and ‘movie’ columns were used as aliases.
- This was done using an inner join which only retrieved matching records from the two datasets
- There were 1244 matching records from the two datasets.
- This new data frame, gross-df, was then joined to the data frame with the top 10 best films using an inner join and the ‘primary title column.
- The resulting data frame was then analyzed to find out the average budgets and profits made from the best-performing films.
- 9 films had matching records with the gross-df data frame.

Definition of terms according to Wikipedia

Production Budget: This budget accounts for the costs incurred during the actual filming process such as equipment rentals, crew salaries, and set design.

The worldwide gross in movies refers to the total revenue generated by a film from all sources worldwide, including ticket sales, merchandise, licensing, and distribution deals. It is a measure of a movie's overall financial success.

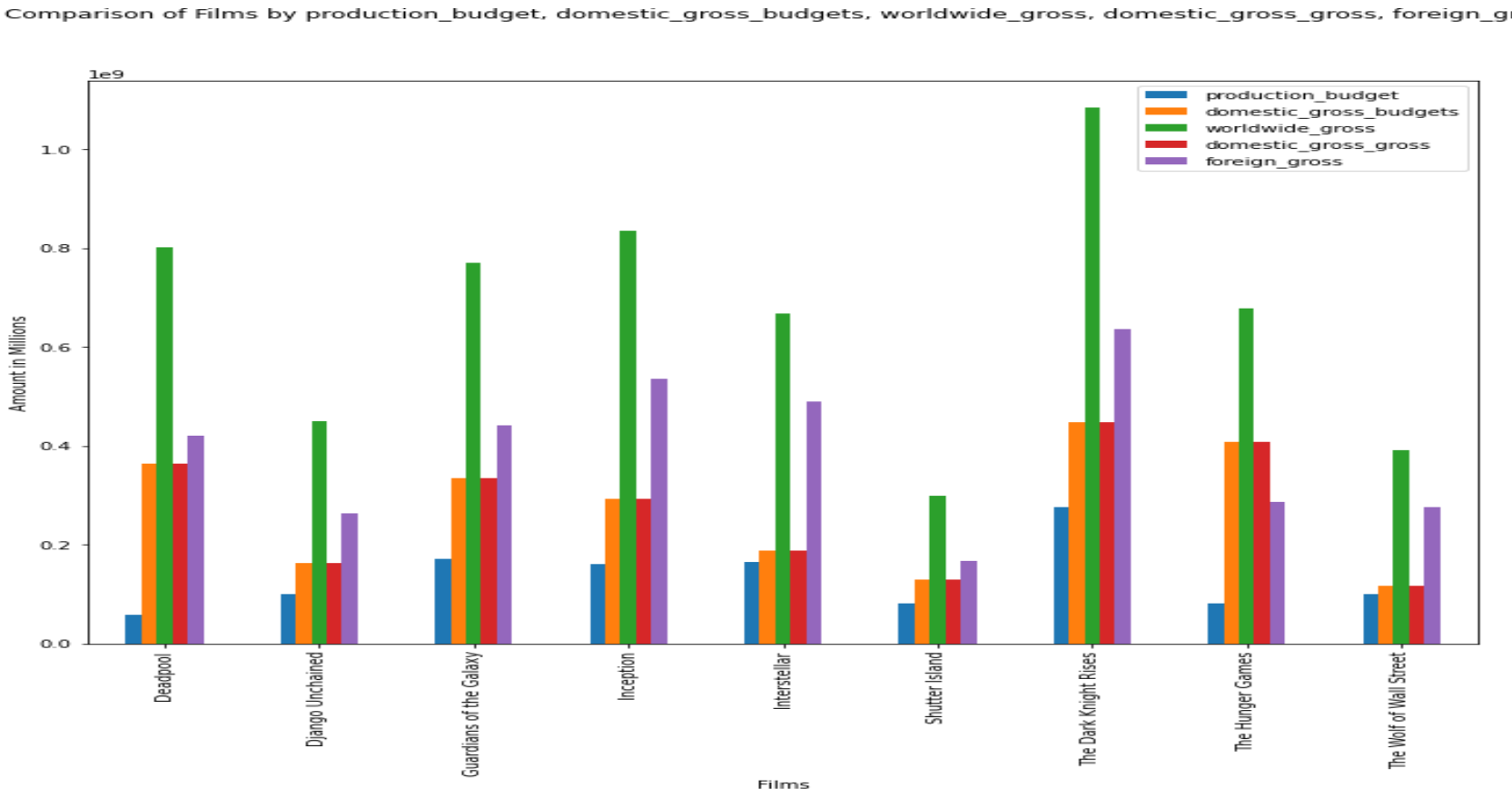
Domestic gross refers to gross box-office revenue from North American countries i.e. U.S., Canada, and Puerto Rico.

Foreign gross includes all other countries in North America.

…findings

- The average budgets and gross earnings made were calculated and the following were the findings:
- The average budget amount in millions of dollars that would be needed to create a great film would be a production budget of 132M.
- The estimated average earnings in million dollars were as follows: worldwide gross of 663M, domestic gross of 271M, and foreign gross of 390M.

- A bar chart showing the budgets and gross earnings of the best-performing films:



CONCLUSIONS AND RECOMMENDATIONS

- In conclusion, this study showed the top 10 best-performing films based on their genres which were Adventure, Action, and Sci-Fi films, the ratings based on the number of viewers' votes, the average budgets needed to create a great film, and the estimated gross earnings from these films.
- My recommendation to Microsoft is to consider the following when creating film content:
- The genres most watched, the ratings from viewers, the budgets and the projected earnings from the films.

THANK YOU