Dataset Analysis Report

Investigate A Dataset

****Note - Used Dataset:** The Movie Database (TMDb)

Link: https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata

I. Introduction:

In this analysis, we explored the TMDb dataset, which contains information on over 10,000 movies. The dataset includes details such as budget, genre, rating, and revenue for each movie. Our goal was to gain insights into various aspects of the dataset, including genre popularity, ratings, revenue, and the impact of factors like budget and audience engagement.

II. Research Questions:

1. Which genres are most popular from year to year?
2. Which genres are most popular each year?
3. Which genres received the highest ratings from year to year?
4. Which genres had the highest revenue/budget from year to year?
5. Do higher budgets lead to higher revenue?
6. What factors impact the increase in revenue?

III. Methods and Analysis:

To investigate these questions, we performed the following steps:

1. Data Loading and Understanding: We loaded the TMDb dataset and conducted initial exploratory analysis to gain an understanding of the data structure and variables.
2. Data Wrangling: during our analysis, we performed necessary data wrangling tasks, including transforming variables, and filtering the dataset based on relevant criteria. We ensured that the data was in a clean and usable format for our analysis.
3. Genre Popularity Analysis: We analyzed the number of votes received by movies in each genre to determine the most popular genres from year to year. We also examined the overall popularity of genres based on the total number of votes.
4. Genre Ratings Analysis: We examined the average ratings of movies in different genres to identify genres that consistently received high ratings. We compared the ratings of popular genres against less popular ones.
5. Revenue and Budget Analysis: We investigated the relationship between revenue and budget across different genres. We explored which genres had the highest revenue and budget figures, highlighting any notable trends or patterns.

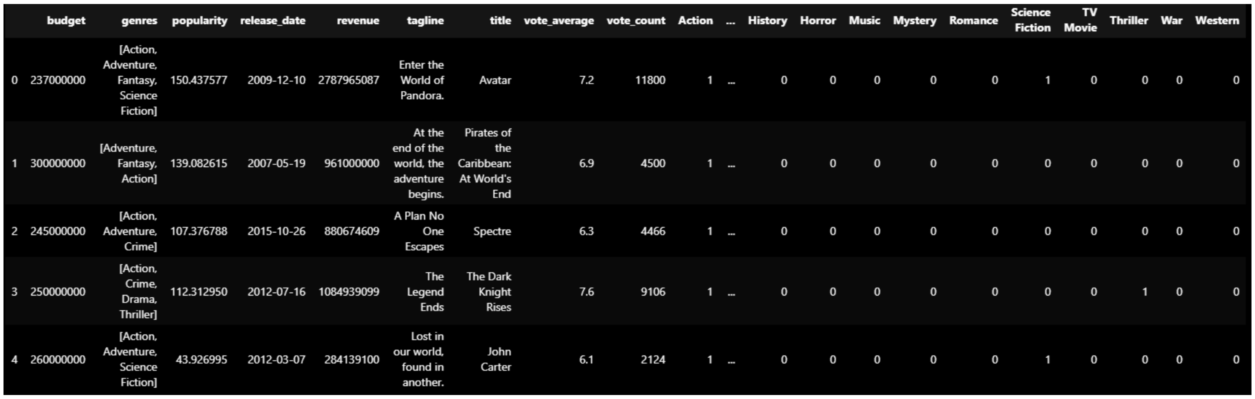
6. Impact of Factors on Revenue: We analyzed the influence of factors like budget, number of votes, and average ratings on movie revenue. We assessed the correlation between budget and revenue and identified genres that had a significant impact on revenue.

IV. Data Wrangling:

During our analysis, we performed necessary data wrangling tasks, including transforming variables, and filtering the dataset based on relevant criteria. We ensured that the data was in a clean and usable format for our analysis.

1. **General Property:**
   We observed data samples, saw the statistic of both text columns and numeric columns and also found some information:
   - Genre Distribution: The genre that appears most frequently in the dataset is "Drama." This indicates that drama movies are highly represented in the collection of over 10,000 movies.
   - Language Composition: An overwhelming majority, approximately 93%, of the movies in the dataset are in English. This suggests that English-language films dominate the dataset.
   - Budget Range: The budgets of the movies in our dataset range from 0 to 380 million dollars. This wide range highlights the varying financial investments made in producing these films.
   - Revenue Range: The revenues generated by the movies in our dataset span from 0 to 2.7 billion dollars. This indicates the potential profitability of movies and the diverse financial success they can achieve.
   - Voting Popularity: The number of votes received for each movie in the dataset ranges from 0 to 13,752 votes. This information provides an insight into the level of audience engagement and interest in these films.
   - These interesting findings add depth and context to our understanding of the movie dataset, and they serve as a foundation for further analysis and exploration.
2. **Data cleaning**
   - Drop certain columns that are unnecessary to answer my question
   - Convert column “genres” to a dictionary column and then split the name of each genre into column using one-hot encoding.
   - Add new column that contain the number of years since the movies were released
   - Add new column that contain the number of votings by year of each movie

   Below is the data sample after I cleaned and transformed to ready to analyze:

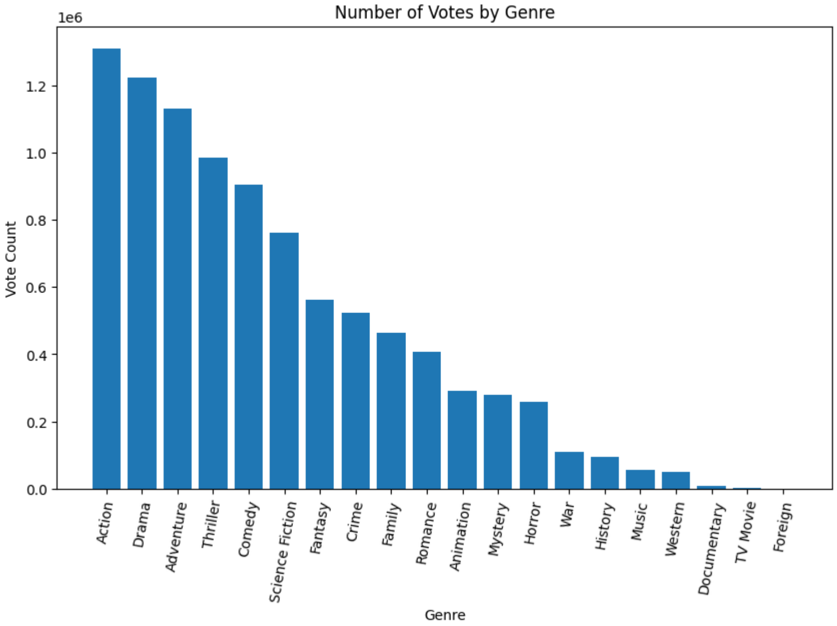


	budget	genres	popularity	release_date	revenue	tagline	title	vote_average	vote_count	Action	...	History	Horror	Music	Mystery	Romance	Science Fiction	TV Movie	Thriller	War	Western
0	237000000	[Action, Adventure, Fantasy, Science Fiction]	150.437577	2009-12-10	2787965087	Enter the World of Pandora.	Avatar	7.2	11800	1	...	0	0	0	0	0	1	0	0	0	0
1	300000000	[Adventure, Fantasy, Action]	139.082615	2007-05-19	961000000	At the end of the world, the adventure begins.	Pirates of the Caribbean: At World's End	6.9	4500	1	...	0	0	0	0	0	0	0	0	0	0
2	245000000	[Action, Adventure, Crime]	107.376788	2015-10-26	880674609	A Plan No One Escapes	Spectre	6.3	4466	1	...	0	0	0	0	0	0	0	0	0	0
3	250000000	[Action, Crime, Drama, Thriller]	112.312950	2012-07-16	1084939099	The Legend Ends	The Dark Knight Rises	7.6	9106	1	...	0	0	0	0	0	0	0	1	0	0
4	260000000	[Action, Adventure, Science Fiction]	43.926995	2012-03-07	284139100	Lost in our world, found in another.	John Carter	6.1	2124	1	...	0	0	0	0	0	1	0	0	0	0

V. Results and Visualizations:

1. Which genres are most popular from year to year?

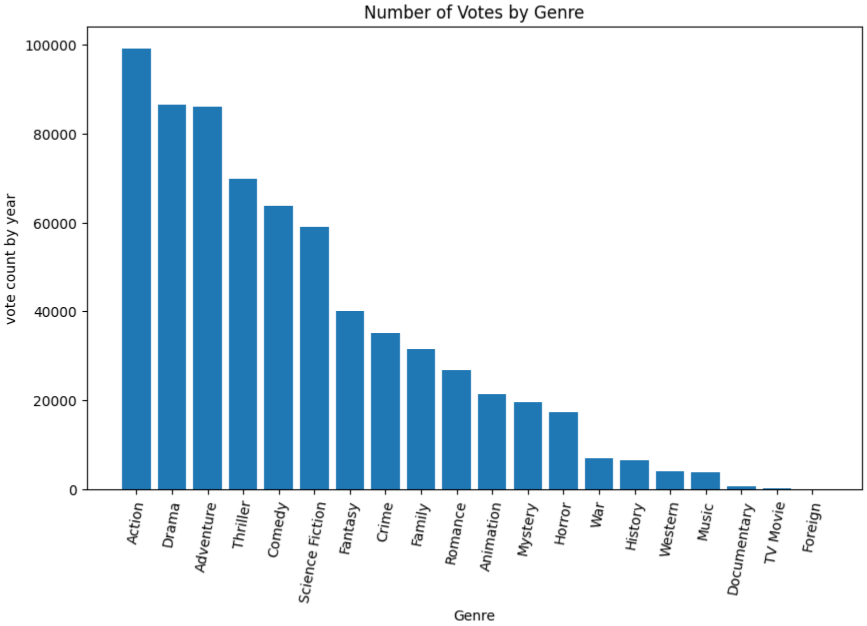
We can observe that Action, Drama, Adventure, Thriller, Comedy and Science Fiction are top 6 genres that most popular from year to year. They also have the number of votings significantly more than other genres. These genres consistently attract a substantial number of votes, indicating their widespread appeal and audience engagement.



On the other hand, War, history, music, westernm documentary and TV movie are the least popular. Movies falling within these genres tend to have fewer than 200,000 votes, suggesting a narrower audience reach and potentially lower levels of overall popularity.

2. Which genres are most popular each year?

In our analysis of the dataset, we have examined the number of votes received by movies in each year. Interestingly, the findings mirror our earlier observations, reinforcing the popularity of certain genres across multiple years. The top six consistently popular genres, known for attracting a substantial number of votes: Action, Drama, Adventure, Thriller, Comedy, Science Fiction.

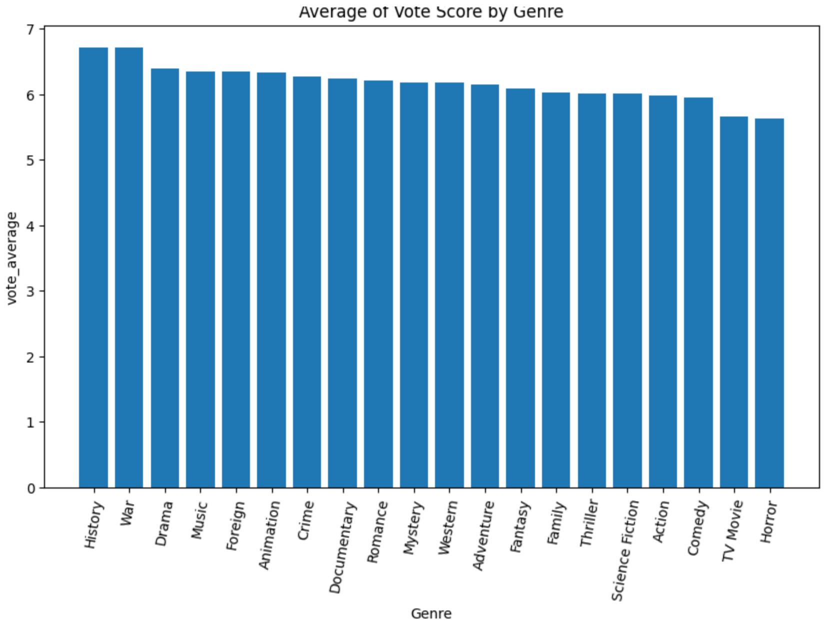


Conversely, we have also noticed that certain genres tend to be less popular, receiving relatively fewer votes over time. These genres include: War, History, Music, Western, Documentary, TV Movie. Movies categorized within these genres generally have a lower number of votes, suggesting a narrower audience reach and potentially lower levels of overall popularity.

These consistent trends in genre popularity across different years offer valuable insights into audience preferences and can assist filmmakers, researchers, and industry professionals in making informed decisions regarding genre selection and target demographics.

3. Which genres received highest review from year to year?

Our analysis uncovers an interesting pattern in the ratings of different genres. Despite receiving fewer votes, genres like History, War, and Music consistently rank among the top genres with the highest ratings. On the other hand, genres like Science Fiction, Action, and Comedy, which are widely popular, tend to have lower average ratings.

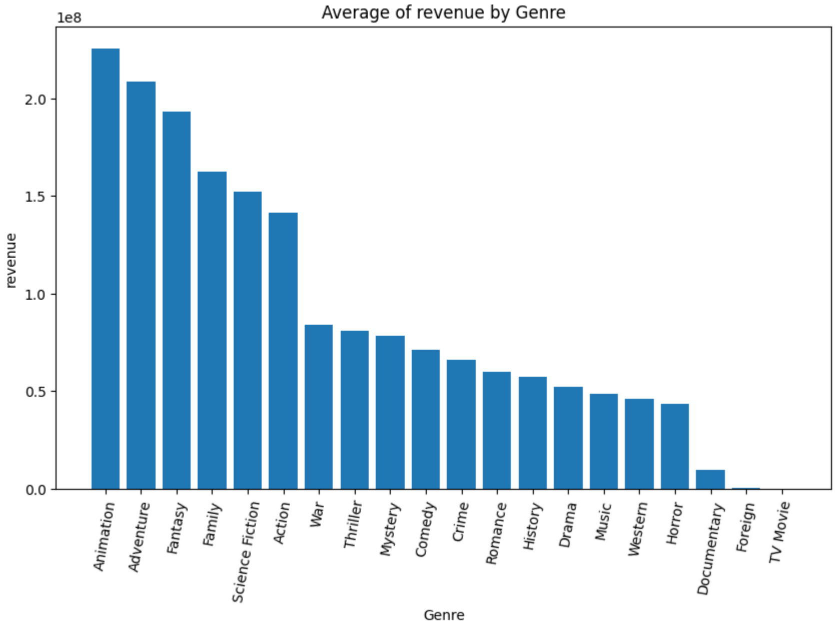


This highlights the distinction between popularity and critical acclaim. Genres with lower vote counts, such as History, War, and Music, often deliver movies that are highly regarded and appreciated by viewers. In contrast, genres with higher vote counts, like Science Fiction, Action, and Comedy, may struggle to consistently achieve high ratings.

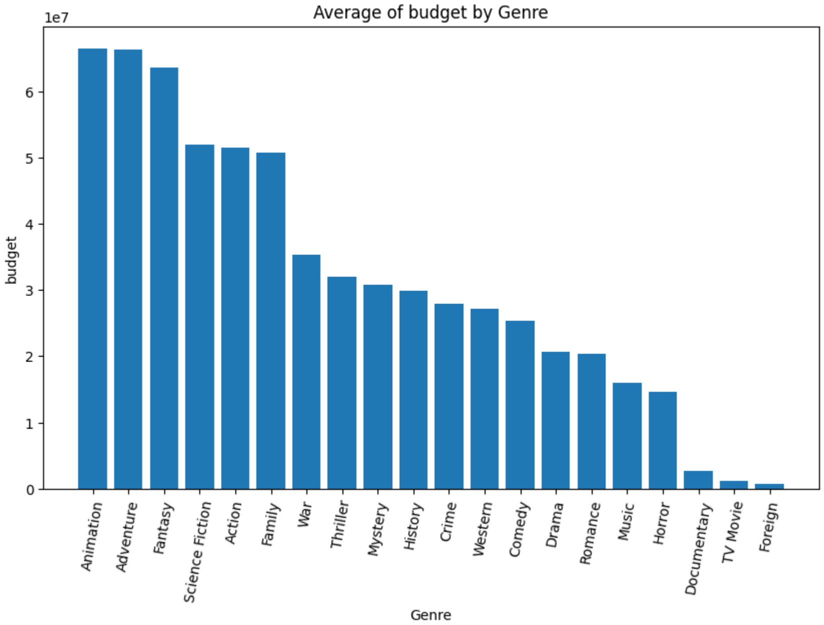
Understanding this difference can help filmmakers and enthusiasts appreciate the diversity of audience preferences and make informed decisions when creating or selecting movies in different genres.

4. Which genres had highest revenue/budget from year to year?

Our analysis reveals a noteworthy relationship between revenue and budget across different genres. Genres like Animation, Adventure, Fantasy, and Science Fiction have the highest revenue, but they also tend to have the highest budgets. Conversely, genres such as Documentary, TV Movie, and Foreign have the lowest revenue and budget.

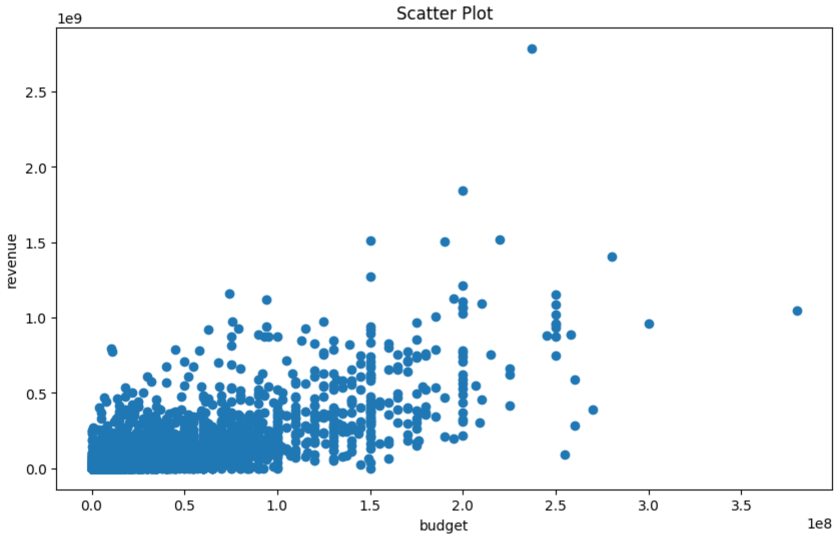


This suggests that higher-budget productions in genres like Animation, Adventure, Fantasy, and Science Fiction have the potential to generate substantial revenue. In contrast, genres with lower budgets, such as Documentary, TV Movie, and Foreign, face challenges in achieving significant financial success.



5. Do the higher budget make the higher revenue?

Upon examining the data, we can observe a modest correlation between revenue and budget. The scatter plot visualization reveals that higher budgets tend to correspond to higher revenues.

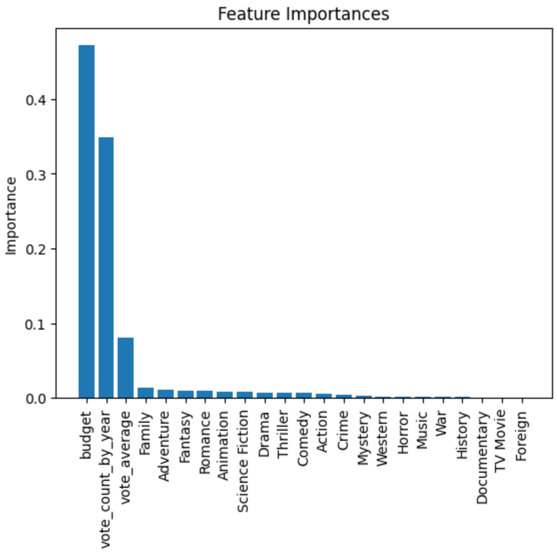


This relationship suggests that movies with larger financial investments have the potential to generate higher revenues. While it is not a definitive rule, the general trend indicates that a higher budget can contribute to increased revenue for a movie.

By recognizing this connection, filmmakers and industry professionals can make informed decisions when allocating resources and estimating potential revenue outcomes for their projects.

6. Which factors impact to the increasing of revenue?

Several factors have been identified as impactful contributors to the increase in movie revenue. Among these factors, the budget has the most significant influence on revenue, followed by the number of votes received per year and the average rating.



The budget allocated to a movie plays a crucial role in its revenue potential. Higher budgets often lead to increased production values, marketing efforts, and overall visibility, which can attract larger audiences and generate higher revenue.

Moreover, certain genres, such as Family, Adventure, Romance, Fantasy, and Animation, have been found to have a particularly strong impact on revenue. Movies within these genres often attract a dedicated fan base and have a higher likelihood of generating significant revenue.

By understanding these influential factors, filmmakers and industry professionals can make informed decisions to optimize revenue generation and maximize the success of their movies.

VI. Conclusion:

1. Results

In this project, we conducted an exploratory analysis of a dataset comprising information on over 10,000 movies from The Movie Database (TMDb). We aimed to gain insights into various aspects of the data, including genre popularity, ratings, revenue, and budget. Here are the key findings:

1. Genre Popularity: Action, Drama, Adventure, Thriller, Comedy, and Science Fiction emerged as the most popular genres consistently over the years. These genres garnered a significant number of votes, indicating widespread appeal and audience engagement. Conversely, War, History, Music, Western, Documentary, and TV Movie genres were found to be less popular, with relatively fewer votes.

2. Genre Ratings: Genres such as History, War, and Music consistently received high ratings, despite their lower popularity. In contrast, popular genres like Science Fiction, Action, and Comedy tended to have lower average ratings. This highlights the distinction between popularity and critical acclaim.

3. Revenue and Budget: Animation, Adventure, Fantasy, and Science Fiction genres had the highest revenue figures, but they also had the highest budgets. Conversely, Documentary, TV Movie, and Foreign genres exhibited the lowest revenue and budget values. This suggests that higher-budget productions in certain genres have the potential to generate significant revenue.

4. Impact on Revenue: Factors influencing revenue include the budget, number of votes received per year, average rating, and genre. Higher budgets generally correlate with higher revenues. Additionally, genres such as Family, Adventure, Romance, Fantasy, and Animation were found to have a stronger impact on revenue.

In conclusion, our analysis provides valuable insights into the dynamics of the movie industry. Understanding genre preferences, the relationship between revenue and budget, and the factors influencing revenue can assist filmmakers and industry professionals in making informed decisions regarding genre selection, resource allocation, and revenue projections for their projects.

2. Limitations

Here are some potential limitations of the data and methods used in this project:

1. Not using all column: In this project we did not use all the column to explore. This may leads to missing some important information or may affect to the quality of the insights we obtained.

2. Missing Data: The dataset might contain missing values for certain variables, such as budget or revenue that we may not process in a suitable way. These missing values could introduce uncertainties and affect the accuracy of the analysis and conclusions drawn. Considerable care should be taken to handle missing data appropriately, such as through imputation or exclusion of incomplete records, to minimize the impact on the results.

3. Causality vs. Correlation: While analyzing the relationships between variables, it's important to note that correlation does not imply causation. Identifying a correlation between factors like budget and revenue does not necessarily mean that one factor directly causes the other. In this project we have not clearly denfine the relationship is correlation or causation.

It is important to be aware of these limitations and provide appropriate context when interpreting and presenting the results of the analysis to ensure a comprehensive and accurate understanding of the findings.