DV Assignment 3

Deepkumar Patel
IMT2021011

Ricky Ashokkumar Ratnani
IMT2021030

Nori Meher Ashish
IMT2021085

December 2023

1 Dataset & Preprocessing

The dataset contains information related to various movies, encompassing a range of details relevant to the film industry. Below is an overview of the key attributes present in the dataset:

- index: A unique identifier or index for each record in the dataset.
- budget: The budget allocated for producing the movie.
- genres: The genres associated with the movie. A single movie may belong to multiple genres.
- homepage: The URL of the movie's homepage.
- id: A unique identifier for each movie.
- keywords: Keywords or phrases associated with the movie, providing insights into its content.
- original_language: The language in which the movie was originally produced.
- original_title: The original title of the movie.
- overview: A brief overview or synopsis of the movie's plot.
- popularity: A measure of the movie's popularity.
- production_companies: Information about the production companies involved in making the movie.
- production_countries: Information about the countries where the movie was produced.
- release_date: The date when the movie was released.
- revenue: The revenue generated by the movie.
- runtime: The duration of the movie in minutes.
- spoken_languages: The languages spoken in the movie.
- status: The release status of the movie (e.g., Released).
- tagline: A tagline associated with the movie, often used for promotional purposes.
- title: The title of the movie.
- vote_average: The average rating given to the movie by viewers.
- vote_count: The number of votes or ratings received by the movie.

- cast: Information about the cast members involved in the movie.
- crew: Information about the crew members involved in the movie.
- director: The director(s) responsible for directing the movie.

The dataset covers a range of genres, languages, and production details, making it suitable for analyses related to movie revenue, genres, and other aspects of the film industry. **We one-hot encoded the genres.**

2 Analysis of Movie Revenue & Genres

2.1 Introduction

The objective of this data story is to explore the relationship between movie revenue and genres. The analysis will involve visualization and data analysis, aiming to uncover patterns and insights that contribute to the understanding of revenue dynamics in the film industry.

2.2 Initial Analysis & Visualisation

2.2.1 Visualization 1: Genre Revenue Distribution

Utilized boxplot visualizations (1) to understand the distribution of revenue for each genre. In the exploration of genre revenue distribution using boxplot visualizations, the primary goal is to gain insights into how movie revenue varies across different genres. Boxplots are effective tools for depicting the spread and central tendencies of revenue within each genre.

Boxplot Analysis

- **Genre Impact on Revenue**: Genres with wider boxplots suggest greater revenue variability, indicating that the choice of genre plays a substantial role in revenue outcomes.
- **Genre Preferences**: We can see that genres like Action, Adventure and Animation have higher median revenue and suggest that audiences prefer these genres.
- **Outlier Detection**: Outliers in certain genres like Action represent movies that significantly outperformed compared to typical revenue expectations. (The data was a little skewed towards the Action genre).
- Even though some of the genres like Thriller and Science Fiction have niche audience can generate high revenues.

Visualisation

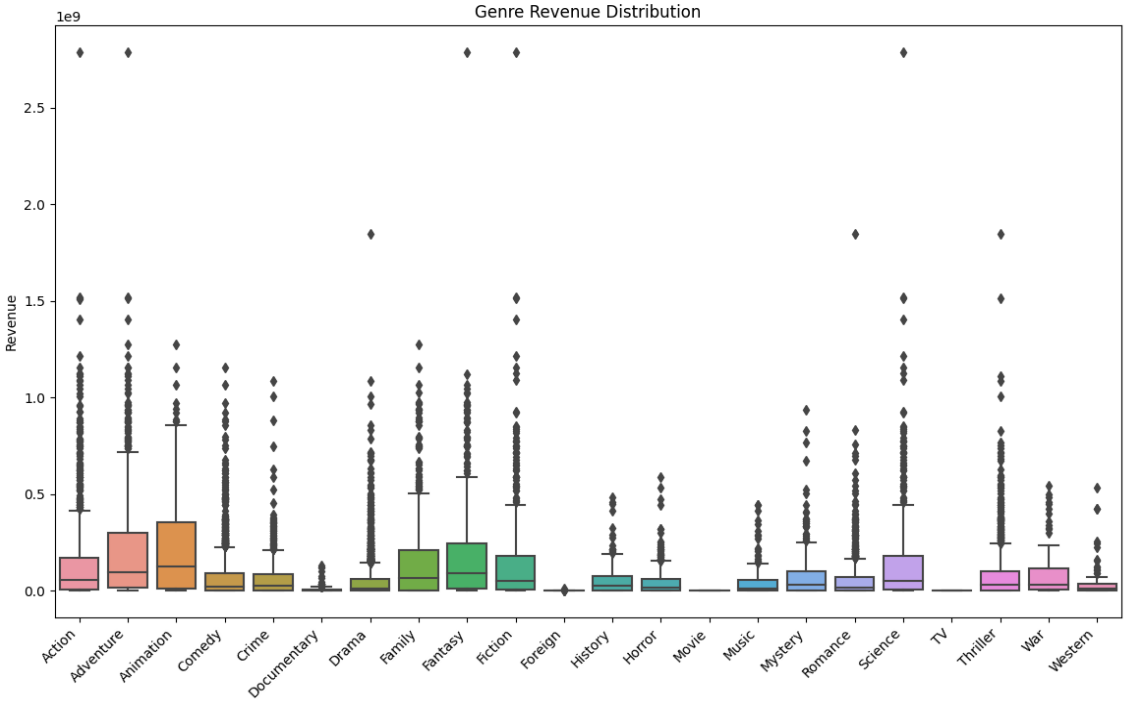


Figure 1: Revenue Distribution

2.2.2 Analytics-1 Genre Revenue Statistics

We can find that from figure 2 genres like Action, Adventure, Science Fiction, Animation have higher revenues indicating that the masses are inclined towards these type of films and the production companies can focus more towards these genres to get more revenues.

2.2.3 Analytics-2 Correlation HeatMap

Analyzing the correlation heatmap **??** between different genres and revenue provides valuable insights into the relationships between genres and movie revenue. Here are some inferences:

- Genres with a positive correlation with revenue suggest that movies belonging to those genres tend to have higher revenues. For Example genres like Adventure and Animation have high correlation it implies that these genres movies are likely to generate higher revenues.
- Identifying Niche Genres, Negative correlations may highlight niche genres that, while not generating high revenues on average, could have a

	Action	Adventure	Fantasy	Science	Fiction	\
mean	2.628476e+08	3.621070e+08	3.329023e+08	2.724090e+08	2.724090e+08	
median	1.484121e+08	2.522769e+08	2.124457e+08	1.265468e+08	1.265468e+08	
std	3.211759e+08	3.546268e+08	3.656091e+08	3.579546e+08	3.579546e+08	
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
max	2.787965e+09	2.787965e+09	2.787965e+09	2.787965e+09	2.787965e+09	

	Crime	Drama	Thriller	Animation	Family	\
mean	1.065425e+08	9.263237e+07	1.312284e+08	3.402887e+08	2.840738e+08	
median	5.198762e+07	3.700000e+07	6.310000e+07	2.972681e+08	1.836497e+08	
std	1.631511e+08	1.587298e+08	2.055788e+08	2.913977e+08	2.891823e+08	
min	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
max	1.084939e+09	1.845034e+09	1.845034e+09	1.274219e+09	1.274219e+09	

	...	Romance	Horror	Mystery	History	\
mean	...	1.021000e+08	7.338487e+07	1.346431e+08	8.208985e+07	
median	...	4.411631e+07	4.026197e+07	6.965378e+07	5.334284e+07	
std	...	1.759473e+08	1.011071e+08	1.756728e+08	8.860729e+07	
min	...	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	
max	...	1.845034e+09	5.853490e+08	9.382127e+08	4.142115e+08	

	War	Music	Documentary	Foreign	TV	Movie
mean	1.177385e+08	9.396434e+07	1.353419e+07	22260.000000	0.0	0.0
median	8.226550e+07	3.272696e+07	2.983065e+05	0.000000	0.0	0.0
std	1.303845e+08	1.261235e+08	2.917640e+07	49774.873179	0.0	0.0
min	1.083480e+05	0.000000e+00	0.000000e+00	0.000000	0.0	0.0
max	5.423074e+08	4.431400e+08	1.273922e+08	111300.000000	0.0	0.0

Figure 2: Central Tendencies for Revenue and Genre

dedicated fan base.

- And we see correlation between the genres i.e an animation movie is strongly correlated with Family genre and a thriller movie is negatively correlated with comedy genre.

2.3 Data Transformation

We create a new column called genre_category indicating whether a movie falls under a a "Single Genre" or "Multiple Genres" category. This transformation simplifies the representation of the genres associated with each movie. In this transformation, the 'genre_category' column is set to "Single Genre" if a movie belongs to only one genre, and "Multiple Genres" otherwise. This categorization provides a simplified view that may make it easier to analyze and interpret relationships between genre categories and revenue.

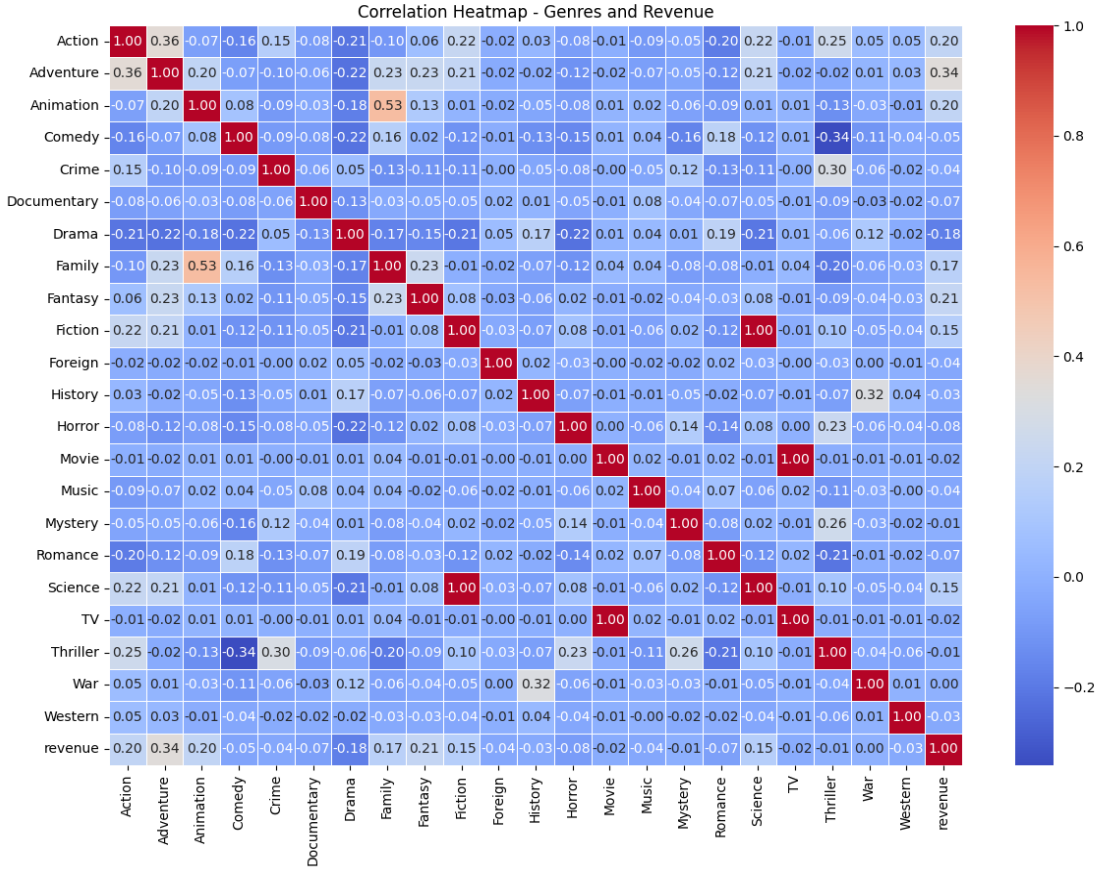


Figure 3: Correlation Matrix between genres and revenue

genre_category	mean	median	count
Multiple Genres	9.209431e+07	23399176.0	3880
Single Genre	4.306632e+07	7600000.0	877

Figure 4: Central Tendencies for Genre Category

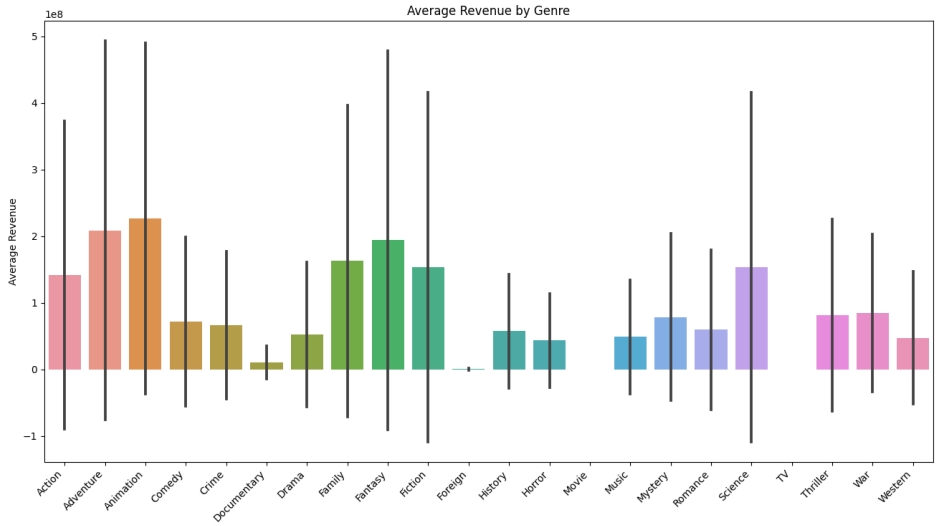


Figure 5: Bar Chart for Average Revenue vs Genre

2.3.1 Analytics-3 Statistical Analysis of Transformed Data

We find the central tendencies for the newly added column genre_category 4. We see that movies of multiple genres have higher mean revenue than single genre movies catering to wider range of audiences. Whereas Single Genre films are catered to specific niche audience who either love that genre or specific aspect of that film like either the cast, director or the crew. Moreover the data is skewed towards the Multiple Genre films so a better dataset which contains equal number of Single and Multiple Genre films might have different results.

2.4 Final Round of Visualization & Analytics

2.4.1 Visualization 2: Average Revenue by Genre

Developed a bar chart to visualize the average revenue for each individual genre.
Inferences

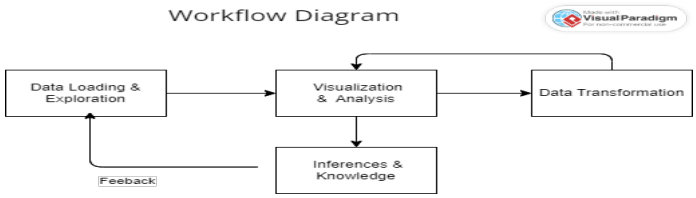


Figure 6: Workflow Diagram

- Analyzing the top-performing genres like Adventure, Animation & Action provides insights into audience preferences and trends.Understanding which genres consistently generate higher revenue can guide filmmakers and studios in targeting specific genres for future productions.
- Genres with lower average revenue could represent opportunities for innovation or untapped audience segments. Exploring ways to enhance the appeal of these genres or experimenting with genre combinations may open up new avenues for success.

Some miscellaneous visualizations for revenue Figure 7 and Figure 8 Note that the highest grosser bar chart for a particular year can be made for any year. I have just shown for 2012 here.

2.5 Feedback Loop

- Identified genres showing strong correlation with high revenue and understood the impact of certain genres on revenue.
- Based on initial inferences, created new features called "Single Genre" and "Multiple genre" to analyse the impact of mutiple genre on revenue.
- Did statistical analysis of newly added features to make out inferences.
- Made final set of visualizations and analytics.

3 Understanding the dependence of popularity of movies on certain factors

3.1 Introduction

In this exploration, we aim to understand the factors influencing movie popularity. We'll be investigating the relationships between popularity and key variables like genre, director, budget, and release date. Our primary focus is on uncovering patterns in how a movie's popularity is connected to its director.

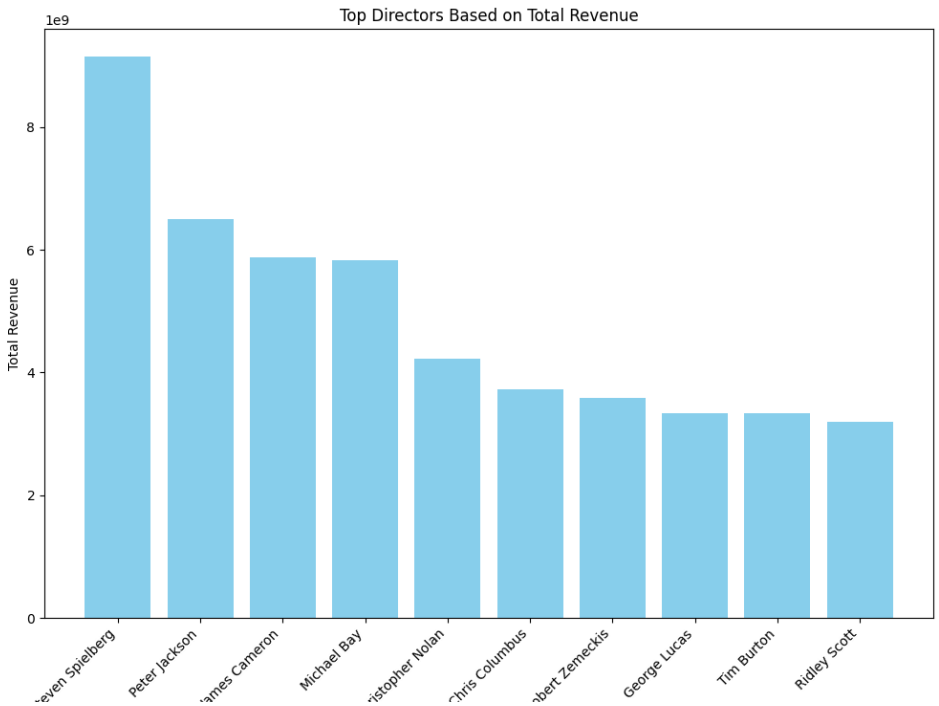


Figure 7: Total Revenue vs Director

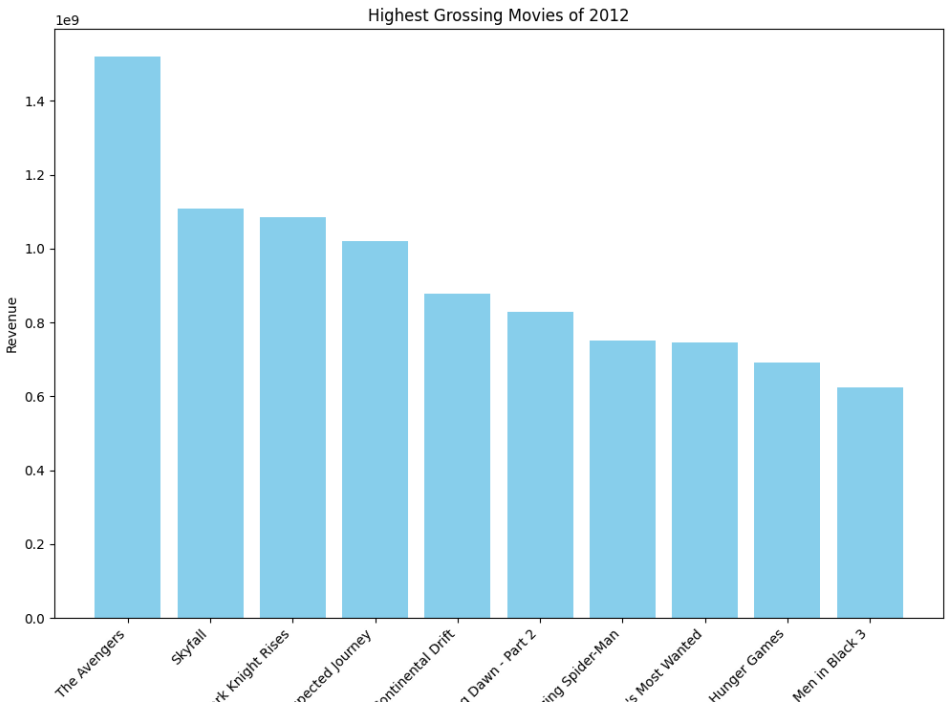
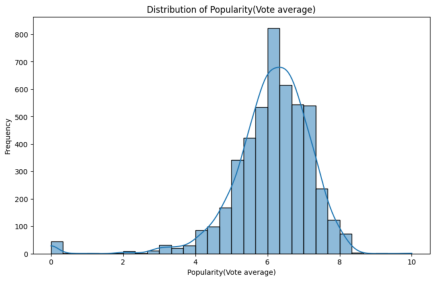


Figure 8: 2012 Highest Grossers



(a) Popularity distribution

mean	6.118071
std	1.123471
min	0.000000
25%	5.600000
50%	6.200000
75%	6.800000
max	10.000000

(b) Stats of popularity distribution

Figure 9: Distribution of popularity and statistical values for it

Popularity measurement : To quantify the popularity of movies, we will utilize the "Vote Average" field. This metric represents the average rating given to a movie by viewers, ranging from 0.00 to 10.00.

3.2 Initial Analysis and preprocessing

- Used the vote average field to calculate a dictionary having the average popularity(vote average) for each director.
- New edges table is create where genre is the source and director is the target. This edge table is used to create the node link diagram in gephi.

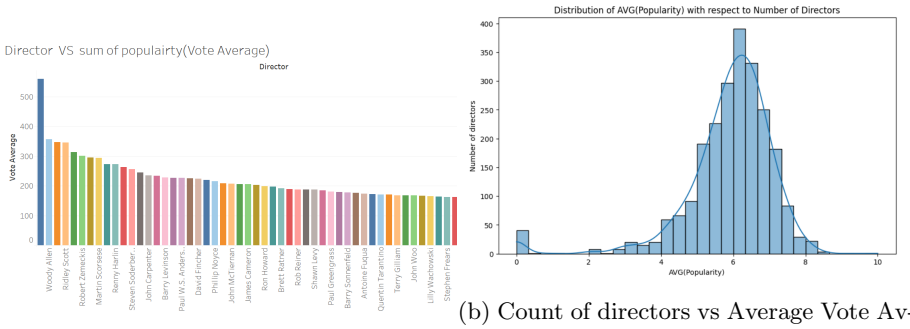
3.3 Statistical Analysis of "Vote Average"

For the basic understanding of how the popularity score is distributed over all the dataset of movies, a hist-plot is being used.

From the Figure 9a we can make these inferences on the basic distribution of popularity,

- From the calculation and above viz we find that the mean popularity of movies is 6.25, with a standard deviation of 1.1234.
  The mean value of 6.11 indicates a moderate level of popularity on average.
  The range is only from 0.00 to 10.00 so the standard deviation of 1.1234(comparatively somewhat small) shows that the popularity scores exhibits some what consistency but still there is some diversity.

- With a median popularity score of 6.20, the dataset's distribution appears a little skewed toward high popularity but otherwise relatively symmetric.



(a) Director vs SUM(Vote Average)

(b) Count of directors vs Average Vote Average

Figure 10: Popularity and director trends

The median, slightly larger than the mean, suggests a balanced distribution of popularity scores. Notably, highly popular movies have some influence on the median.

These insights provides a nuanced understanding of the distribution and variability of popularity scores, laying the foundation for more in-depth explorations into the factors influencing movie popularity.

3.4 Initial Analysis of Dependence of Popularity on Director

Our initial assumption posits that a movie's director significantly influences its popularity. To explore this, we first examine trends in the sum of "Vote Average" for all movies with respect to each director.

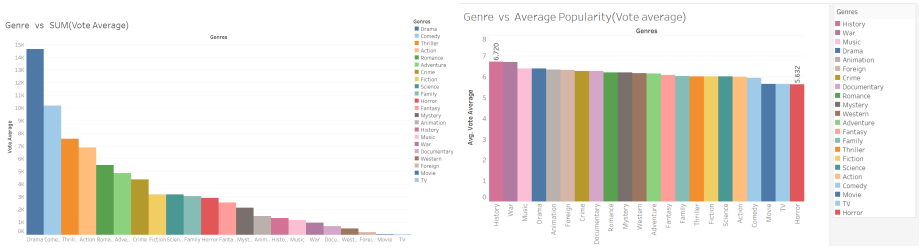
3.4.1 Director Impact on Popularity

The visualization in Figure 10a illustrates the sum of "Vote Average" for each director, providing insights into directors with higher cumulative popularity(This includes both the facts that a director created more number of movies and movies with more popularities).
A clear decreasing trend is visible in this suggesting that the cummulative popularity indeed varies based on the director of the movie.

3.4.2 Director-Specific Considerations

We can see the trend in cummulative value but for getting the true idea of whether there is an actual trend or not we should analyse the average popularity trend to remove the influence of the number of movies created by the particular director.

- We can see from Figure 10b that after taking the average of popularity most of the directors have the values between 5.4 to 6.6. This shows



(a) Genre vs SUM(Vote Average) (b) Genre vs Average Vote Average

Figure 11: Popularity and genre trends

that the popularity varies because of the directors but not drastically as the standard deviation is 1.253. Also note that the current distribution properly forms the basis of the popularity distribution we saw earlier.

3.5 Analysis of Dependence of Popularity on Genre

Our initial assumption posits that a movie's genre significantly influences its popularity. To explore this, we examine trends in the sum of "Vote Average" for all movies within each genre.

3.5.1 Genre Impact on Popularity

The visualization in Figure 11a illustrates the sum of "Vote Average" for each genre, providing insights into genres with higher cumulative popularity. This visualization aids in understanding which genres tend to produce more popular movies based on viewer ratings. For example from this we can say that genres like drama, comedy, thriller are much more popular in viewers cummulatively i.e that there are many movies in these genre and they also have good popularity among viewers.

3.5.2 Genre-Specific Considerations

While certain genres may exhibit higher cumulative popularity, for studying the measure of the genre itself it's crucial to consider other factors such as the number of movies in each genre. A high cumulative popularity might be influenced by a large number of movies within that genre. To account for this, we further study the genre versus the average "Vote Average."

From the visualization in Figure 11b, we observe that, while there may be observable discrepancies in the sum of popularity across genres, the average popularity for each genre is relatively consistent. The average popularity ranges from 6.720 to 5.632 across 22 genres. Overall, we can infer that some genres are generally preferred by viewers more than others on average, but the bias from genre is not significantly pronounced.

3.5.3 Genre-Based Popularity Trends Over Time

The dataset spans movies released from 1940 to 2008, encompassing a wide time range. The trends discussed above consider this entire time span. However, different time periods may exhibit distinct trends. Merging all the trends yielded the insights discussed above, considering movies from various time periods.

3.6 Improving the initial visualization :-

From the above analysis of genre vs popularity we saw that genre does effect the popularity. So it can happen that some directors only creates movies of certain genres only and so as those genres being more popular in viewers the director gets more votes. So basically genre might have some bias in the popularity score of the director. So we will give weights to the movies of the directors based on the genre to negate the bias.

3.6.1 Bias by the director and tackling it

In the Figure 12 part (a) the big colored circles represents different genres and the tiny circles are the directors. From the figure we can see that most of the directors tend to create their movies in 2 to 3 distinct field of genre. So each movie of that director being in these 2 genres and those genre being popular or unpopular gives rise to bias induced by genres.

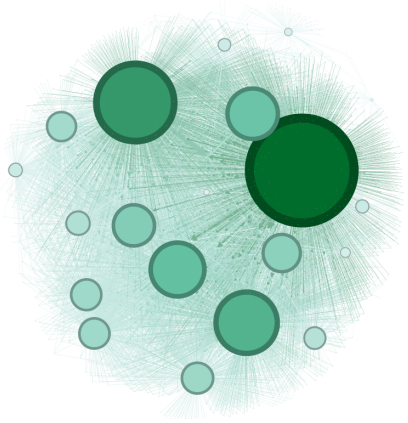
So for tackling it we give more weight to the genres which has less average popularity and vice versa. By doing so, the average popularity calculated for the directors now will not have any bias created by some genres being more popular than other.

Figure 14 shows the visualization for the count of directors wrt average popularity after doing the bias correction. Note that the values are not scaled between 0 to 10 now. So we will only comment on the standard deviation which is mostly the value of interest.

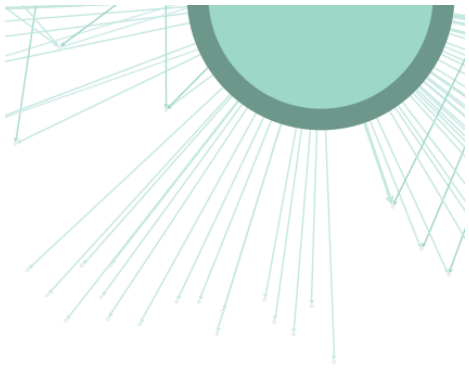
From the Figure 14 we can see that still the trend is almost the same but the plot is more skewed towards more popular than before and the standard deviation is more than before. So we can say that bias from genres before was causing the average popularity to converge to a moderate value at the center.

3.7 Improvements and feedbacks taken

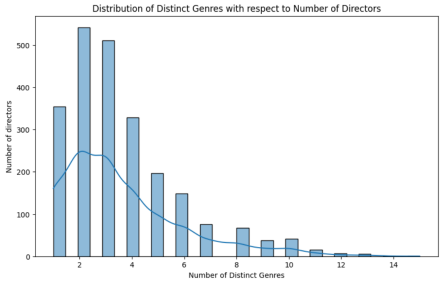
- Concluded that some genres being more popular created bias in other studies.
- Did bias correction using new compuatated column containing weighted popularity.
- Made new inferences using the bias corrected visualizations.



(a) Nodelink diagram showing network of genres and directors.



(b) Director nodes which are only connected to one genre showing the bias.



(c) Histogram showing directors preferring to work in small number of genres

Figure 12: Bias by directors

```
genre_weights
✓ 0.0s
{'History': 0.41504510908260267,
 'War': 0.4153982359011351,
 'Music': 0.4359338994373085,
 'Drama': 0.4363257950634851,
 'Animation': 0.4399142935015862,
 'Foreign': 0.4402989813385685,
 'Crime': 0.4438875395351204,
 'Documentary': 0.44466199715231974,
 'Romance': 0.44958044358224064,
 'Mystery': 0.44982078288869015,
 'Western': 0.45166297326582905,
 'Adventure': 0.4529862046328188,
 'Fantasy': 0.4577008690059655,
 'Family': 0.4621875755281804,
 'Thriller': 0.46317963376566995,
 'Fiction': 0.4633038081004319,
 'Science': 0.4633038081004319,
 'Action': 0.4651328253975723,
 'Comedy': 0.46935210919014697,
 'Movie': 0.492541962016158,
 'TV': 0.492541962016158,
 'Horror': 0.49523919149758133}
```

Figure 13: Wights given to each genre to remove the bias

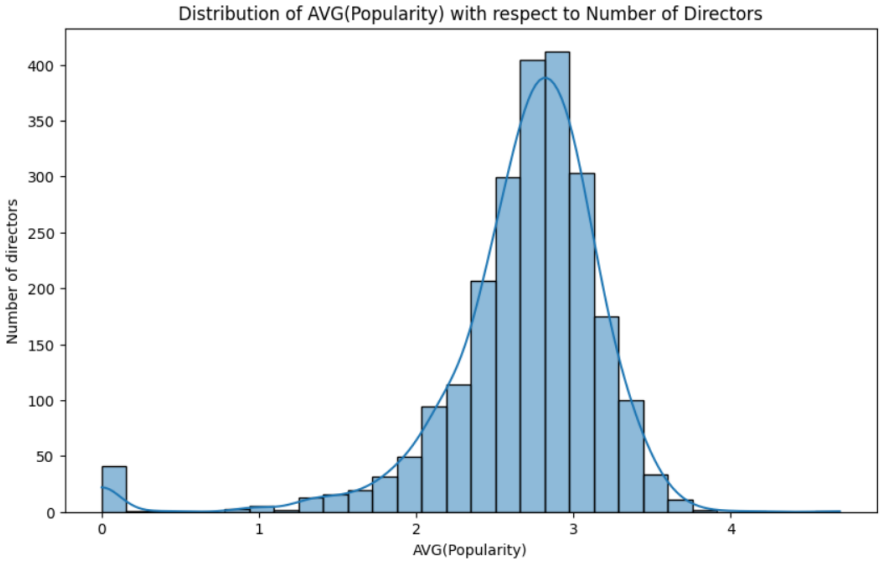


Figure 14: Distribution of directors after bias correction

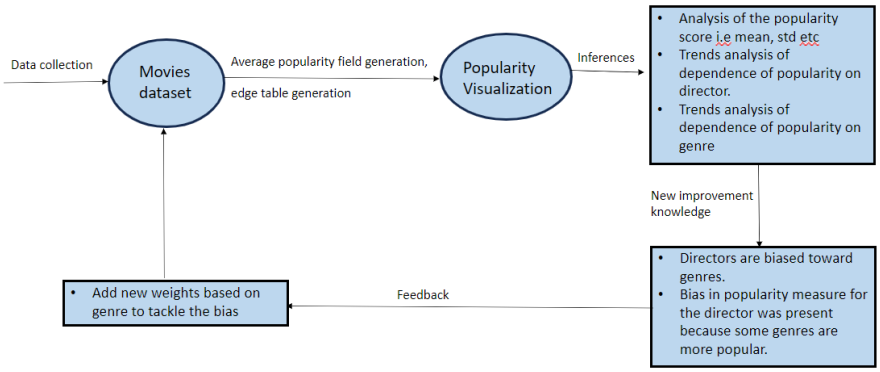


Figure 15: Workflow diagram as per the popularity analysis story

4 Budget

In filmmaking, the budget stands as a linchpin, quietly dictating the course of artistic vision and cinematic brilliance. Far more than a financial outline, a movie's budget is the silent force steering the ship of creativity, impacting everything from visual effects to narrative depth. and hence we take a deep dive into understanding the budget statistics.

4.1 Initial Analysis and Preprocessing

In preprocessing, we create a binary column for each genre using python lists, we also drop some columns that are not essential for our analysis like the main website page for the film. the preprocessing is attached in the jupyter notebook. When it comes to statistics, our fundamentals, like mean are valuable. they can be found in the following table 1. Originally, we found some of the data at 0 budget, which we know is an outlier, so we removed those rows in our analysis, we applied a filter that the budgets should be a minimum of 1e5. One can find the budget histogram and boxplot in figure 16a and 16b. It is quite evident that most of the films have a budget of less than 1e8.

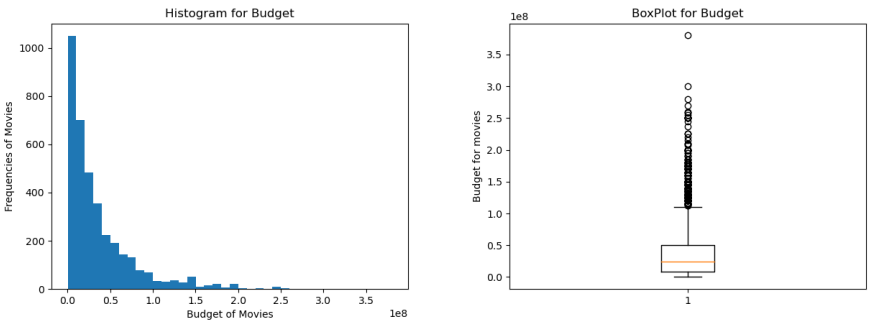
for a better understanding of the distribution of budgets, we plot a histogram and a boxplot.

4.2 Genre and Budget

Genre in movies serves as a guiding force, shaping narratives and influencing audience expectations. It's a crucial aspect that not only defines a film's essence but also significantly impacts budgets. The choice of genre intricately interweaves artistic vision with financial considerations, playing a pivotal role in the success and financial viability of a production.

Metric	Value
count	3.702000e+03
mean	3.767909e+07
std	4.273560e+07
min	1.000000e+05
25%	9.000000e+06
50%	2.400000e+07
75%	5.000000e+07
max	3.800000e+08

Table 1: fundamental statistics for budget



(a) Histogram for Budget (b) BoxPlot for Budget

Figure 16: Some Budget Visualizations

Genre	Mean	Std	Median
Action	58980213.745	55246680.030	40000000.000
Music	22022330.414	21291476.462	15000000.000
Thriller	38209139.667	38063547.876	26000000.000
Romance	27012370.626	26325483.977	19000000.000
Western	34585212.516	47402177.148	17000000.000
Mystery	36703974.055	33649461.752	27000000.000
War	38609875.000	36954087.670	25000000.000
Animation	75834247.688	53417132.868	70000000.000
History	35487848.982	33775340.908	25000000.000
Adventure	73079805.344	62032454.161	55000000.000
Crime	32964513.485	31925529.617	25000000.000
Foreign	2425000.000	2171261.154	1400000.000
Science	59300894.316	58424445.699	40000000.000
Family	63710048.502	52276818.742	50000000.000
Fantasy	72186510.397	62438466.783	55000000.000
Documentary	5951845.000	10372655.429	2000000.000
TV	3066666.667	1792577.288	4000000.000
Comedy	33816508.843	33822856.669	24000000.000
Drama	27839532.012	30438703.581	18000000.000
Horror	19493369.054	23299537.529	12000000.000

Table 2: fundamental statistics for the budget when grouped by genres.

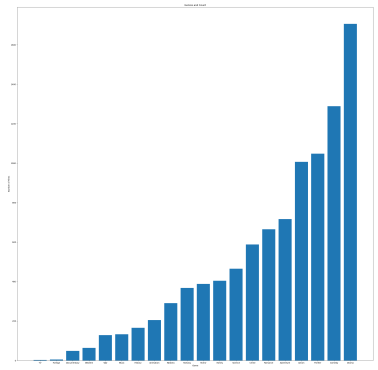
We know a movie can have multiple genres, so here we look at the correlation matrix in figure 17b, which serves as a measure of which genres go together at the box office. The correlation between crime and thriller is 0.31, which means there is a higher chance that both of them come together unlike TV and mystery which correlate at 0.01.

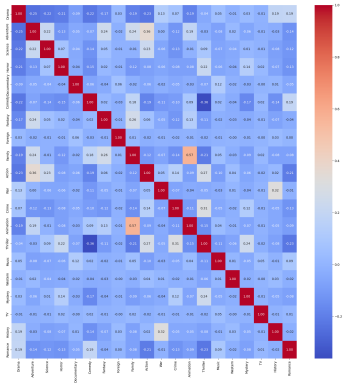
Apart from the correlation, let's look at table 2, for the fundamental statistics for various genres. let's also look at the boxplots and histograms of the various genres in figures 17a and 18 to determine where the budget is going.

looking at the bar chart 17a, it is quite evident that Action, Comedy, and Drama are the most popular genres in movies. Genres like TV and Documentary are not quite made as frequently as the previously mentioned genres.

4.3 Companies and Budget

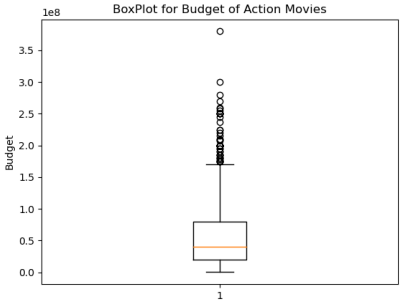
In the dynamic landscape of filmmaking, production companies wield immense influence, steering the course of cinematic ventures. As architects of creativity and financial stewards, these entities play a pivotal role in determining movie budgets and orchestrating the delicate balance between artistic ambition and financial viability. This exploration delves into the intricate interplay between production companies and movie budgets, shedding light on the mechanisms that shape the silver screen's most captivating narratives.

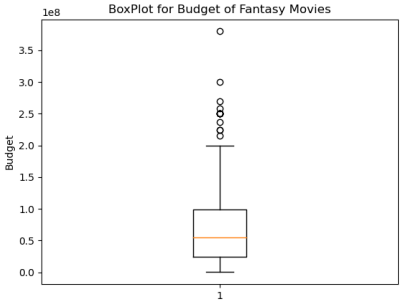




(a) Genres and Movie Counts (b) Correlation matrix among Genres

Figure 17: Visualizations for genres





(a) BoxPlot for Budget of Action Films (b) BoxPlot for Budget of Fantasy Films

Figure 18: BoxPlots for budget of various genres of film. all of them can be found in the files.

Production Company	Mean	Std	Median	# films
Warner Bros.	26592521.746	39767118.579	13000000.000	268
Universal Pictures	28775309.242	46396891.625	15000000.000	240
Paramount Pictures	28744889.792	40827342.667	15000000.000	212
Twentieth Century Fox	30344906.039	40861751.805	15000000.000	180
Columbia Pictures	33175023.986	50732441.655	13000000.000	143
New Line Cinema	35055030.832	49655931.727	15000000.000	119
Relativity Media	21895857.143	32689671.201	12000000.000	98
Metro-Goldwyn-Mayer	31830506.159	42740387.515	15000000.000	88
Touchstone Pictures	34890126.024	51951259.353	12000000.000	83
Columbia Pictures	27658841.139	37187313.948	15000000.000	79

Table 3: fundamental statistics for the budget when grouped by production companies. showing the top 10 here.

We start by looking at the number of movies each company made in the dataset, then look at the average budget of the movies they make. for simplicity, we give the top 10 companies (with the most films) in the table **??**,

When we talk about companies, we also like to talk about what genres the companies are more invested in so we give an example for Warner Bros. in the following pie chart 20a. looking at the pie chart it is quite evident that Drama, Comedy, Action, and Thriller make more than 50% of their films. They are less interested in genres like Documentary which takes up less than 1%.

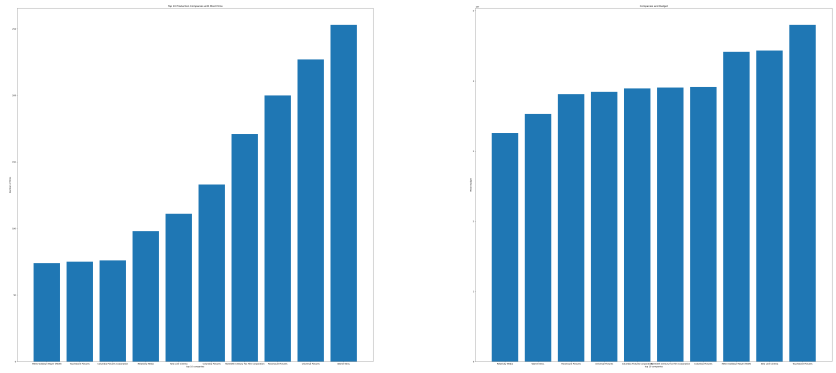
4.4 Directors and Budget

Directors, the visionary architects of cinema, hold the power to transform stories into visual masterpieces. Central to their creative prowess is the management of movie budgets, a skill that impacts the scale and quality of their cinematic endeavors. Examining this dynamic relationship, we find exemplars like Christopher Nolan, whose meticulous approach to budgeting has enabled the realization of groundbreaking films, showcasing the intersection of artistic brilliance and financial insight in the world of filmmaking.

similar to companies, we also talk about what genres the directors are more invested in so we give an example of a crowd favourite, Christopher Nolan. In the Dataset, Christopher Nolan has directed 8 films. The budget statistics for the same can be seen in table 4.

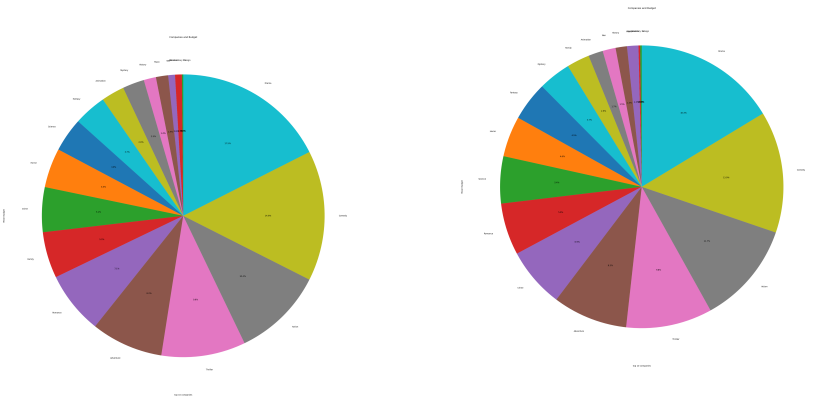
5 Contributions

Preprocessing was done by everyone.
Meher Ashish Nori (IMT2021085) - **Analysis of Movie Revenue & Genres Data Story.**
Ricky Ratnani (IMT2021030) - **Analysis of the dependence of popularity**



(a) Number of films from top 10 companies (b) Mean Budget of top 10 companies

Figure 19: Viz for top 10 companies

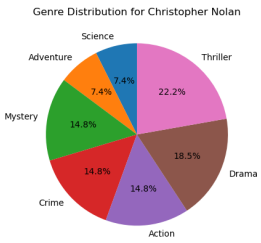


(a) Warner Bros. distribution of genres (b) Paramount Pictures distribution of genres

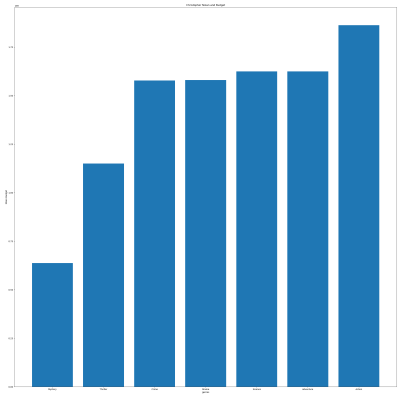
Figure 20: distribution of genres of various companies

Metric	Value
count	8.000000e+00
mean	1.256250e+08
std	8.417149e+07
min	9.000000e+06
25%	4.450000e+07
50%	1.550000e+08
75%	1.700000e+08
max	2.500000e+08

Table 4: Fundamental Statistics for the films directed by Christopher Nolan.



(a) the pie chart of various genres of interest of Chirstopher Nolan



(b) the bar chart for the mean budget films, when grouped by genres, of interest of Christopher Nolan, I apologize if the image is not clear.

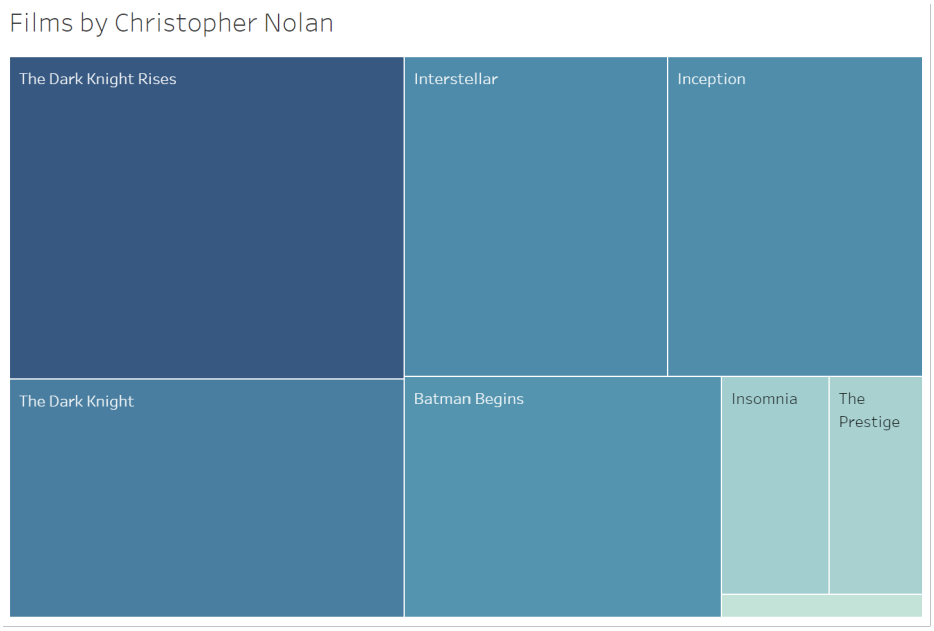


Figure 22: Treemap of the Budget of the Films by Christopher Nolan. Note one can make treemaps on the basis of Genres and Production companies, it was just too cluttered for us to include it here.

of movies on different factors.
Deep Kumar (IMT2021011) - **Budget**