TOP IMDB MOVIES ANALYSIS PROJECT REPORT

(Project Semester January-May 2024)

TOP IMDB MOVIES ANALYSIS

Submitted by

Pratyush Singh

Registration No: 12106287

Programme and Section: K21DB

Course Code: INTB233

Under the Guidance of

Baljinder Kaur , UId: 27952

Discipline of CSE/IT

Lovely School of Computer Science And Engineering

Lovely Professional University, Phagwara

LOVELY PROFESSIONAL UNIVERSITY

Transforming Education Transforming India

CERTIFICATE

This is to certify that Pratyush Singh bearing Registration no. 12106287 has completed INTB233 project titled, **“***TOP IMDB MOVIES ANALYSIS***”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Baljinder Kaur UId-27952
School of Computer Science And Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 18/04/2024

DECLARATION

I, Pratyush Singh, student of BTech Computer Science under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date:18/04/2024 Signature

Registration No. 12106287 Pratyush Singh

ACKNOWLEDGEMENT

I would like to express my sincere thanks to Lovely Professional University for providing the necessary resources and facilities that facilitated this project.

I am grateful to my colleagues and friends for their encouragement and insightful discussions, which greatly contributed to the development of this report.

Lastly, I deeply appreciate the unwavering support of my family during this endeavour.

Pratyush Singh

Date:18/04/2024

INTRODUCTION

The world of cinema is vast, encompassing diverse genres, storytelling styles, and audience preferences. The Internet Movie Database (IMDb) has become a central repository of film information, including user ratings, critical scores, box office figures, and more. By analyzing an extensive dataset from this expansive resource, we can unlock valuable insights into trends within the film industry, audience preferences, and the factors that correlate with both commercial and critical success. This project delves into a rich IMDb dataset, using visualizations and statistical techniques to uncover hidden patterns and potential stories contained within the data.

OBJECTIVE

The primary objective of this analysis is to illuminate the complex dynamics governing filmmaking, reception, and audience engagement. Specifically, this project aims to answer the following key questions:

Audience Preferences: Which genres consistently resonate with IMDb users, as reflected in their ratings? Are there underappreciated categories or niche audiences that emerge?

Critical Reception: Do professional critic scores (such as Meta scores) align with audience sentiment? Are some genres more critically favoured than others?

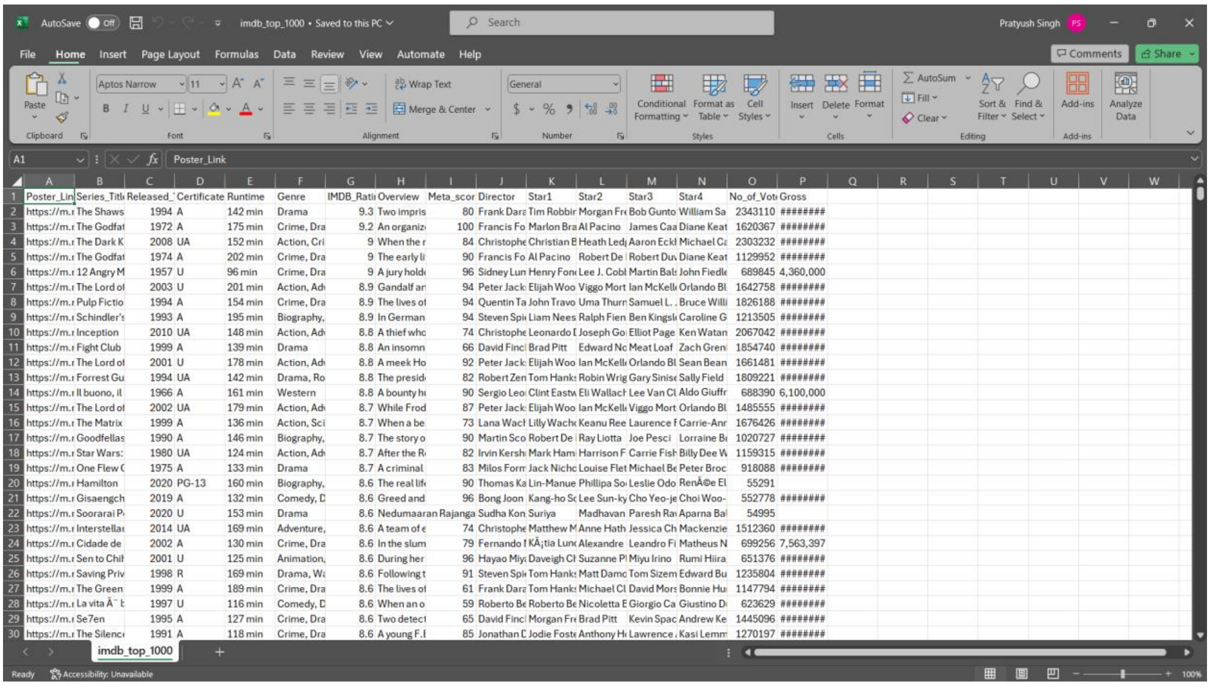
Financial Performance: Can we identify any predictors of box office success? Do factors like rating scores, genre, or release year correlate with a film's gross revenue?

Trends over Time: How has film production volume changed over the decades? Are there notable peaks and valleys in release patterns, and do they tie into broader cultural or industry shifts?

Unexpected Discoveries: Are there anomalies or surprising results that challenge preconceptions about movies?

By meticulously exploring these questions, this analysis will shed light on audience tastes, artistic trends, and commercial success within the captivating world of cinema.

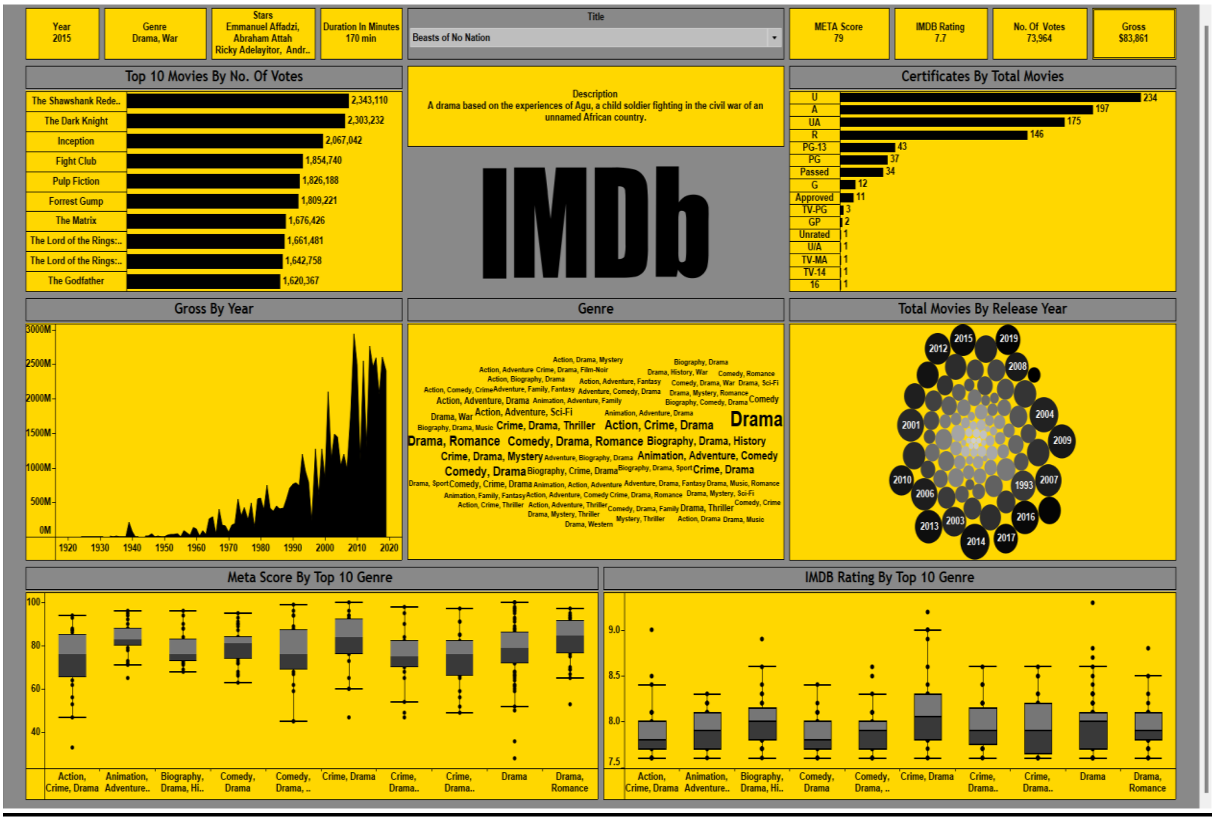
SOURCE OF DATASET



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Poster_Lin	Series_Titl	Released_'	Certificate	Runtime	Genre	IMDB_Ratii	Overview	Meta_scor	Director	Star1	Star2	Star3	Star4	No_of_Vot	Gross
2	https://m.r	The Shaws	1994	A	142 min	Drama	9.3	Two impris	80	Frank Dara	Tim Robbir	Morgan Fre	Bob Gunto	William Sa	2343110	########
3	https://m.r	The Godfat	1972	A	175 min	Crime, Dra	9.2	An organiz	100	Francis Fo	Marlon Bra	Al Pacino	James Caa	Diane Keat	1620367	########
4	https://m.r	The Dark K	2008	UA	152 min	Action, Cri	9	When the r	84	Christophe	Christian B	Heath Led	Aaron Eckl	Michael Ca	2303232	########
5	https://m.r	The Godfat	1974	A	202 min	Crime, Dra	9	The early li	90	Francis Fo	Al Pacino	Robert De	Robert Duv	Diane Keat	1129952	########
6	https://m.r	12 Angry M	1957	U	96 min	Crime, Dra	9	A jury hold	96	Sidney Lun	Henry Fon	Lee J. Cobl	Martin Bals	John Fiedle	689845	4,360,000
7	https://m.r	The Lord of	2003	U	201 min	Action, Adv	8.9	Gandalf an	94	Peter Jack	Elijah Woo	Viggo Mort	Ian McKell	Orlando Bl	1642758	########
8	https://m.r	Pulp Fictio	1994	A	154 min	Crime, Dra	8.9	The lives of	94	Quentin Ta	John Travo	Uma Thurn	Samuel L. .	Bruce Willi	1826188	########
9	https://m.r	Schindler's	1993	A	195 min	Biography,	8.9	In German	94	Steven Spi	Liam Nees	Ralph Fien	Ben Kingsl	Caroline G	1213505	########
10	https://m.r	Inception	2010	UA	148 min	Action, Adv	8.8	A thief who	74	Christophe	Leonardo D	Joseph Go	Elliot Page	Ken Watan	2067042	########
11	https://m.r	Fight Club	1999	A	139 min	Drama	8.8	An insomn	66	David Finc	Brad Pitt	Edward No	Meat Loaf	Zach Gren	1854740	########
12	https://m.r	The Lord of	2001	U	178 min	Action, Adv	8.8	A meek Ho	92	Peter Jack	Elijah Woo	Ian McKell	Orlando Bl	Sean Bean	1661481	########
13	https://m.r	Forrest Gu	1994	UA	142 min	Drama, Ro	8.8	The presid	82	Robert Zen	Tom Hanks	Robin Wrig	Gary Sinise	Sally Field	1809221	########
14	https://m.r	Il buono, il	1966	A	161 min	Western	8.8	A bounty h	90	Sergio Leo	Clint Eastw	Eli Wallach	Lee Van Cl	Aldo Giuffr	688390	6,100,000
15	https://m.r	The Lord of	2002	UA	179 min	Action, Adv	8.7	While Frod	87	Peter Jack	Elijah Woo	Ian McKell	Viggo Mort	Orlando Bl	1485555	########
16	https://m.r	The Matrix	1999	A	136 min	Action, Sci	8.7	When a be	73	Lana Wach	Lilly Wach	Keanu Ree	Laurence F	Carrie-Ann	1676426	########
17	https://m.r	Goodfellas	1990	A	146 min	Biography,	8.7	The story o	90	Martin Sco	Robert De	Ray Liotta	Joe Pesci	Lorraine Br	1020727	########
18	https://m.r	Star Wars:	1980	UA	124 min	Action, Adv	8.7	After the R	82	Irvin Kersh	Mark Ham	Harrison F	Carrie Fish	Billy Dee W	1159315	########
19	https://m.r	One Flew C	1975	A	133 min	Drama	8.7	A criminal	83	Milos Form	Jack Nicho	Louise Flet	Michael Be	Peter Broc	918088	########
20	https://m.r	Hamilton	2020	PG-13	160 min	Biography,	8.6	The real life	90	Thomas Ka	Lin-Manue	Phillipa So	Leslie Odo	RenÃ©e El	55291	
21	https://m.r	Gisaengch	2019	A	132 min	Comedy, D	8.6	Greed and	96	Bong Joon	Kang-ho Sc	Lee Sun-ky	Cho Yeo-je	Choi Woo-	552778	########
22	https://m.r	Soorarai P	2020	U	153 min	Drama	8.6	Nedumaaran Rajanga		Sudha Kon	Suriya	Madhavan	Paresh Ra	Aparna Bal	54995	
23	https://m.r	Interstellar	2014	UA	169 min	Adventure,	8.6	A team of e	74	Christophe	Matthew M	Anne Hath	Jessica Ch	Mackenzie	1512360	########
24	https://m.r	Cidade de	2002	A	130 min	Crime, Dra	8.6	In the slum	79	Fernando M	KÃ¡tia Lund	Alexandre	Leandro Fi	Matheus N	699256	7,563,397
25	https://m.r	Sen to Chih	2001	U	125 min	Animation,	8.6	During her	96	Hayao Miya	Daveigh Ch	Suzanne Pl	Miyu Irino	Rumi Hiira	651376	########
26	https://m.r	Saving Priv	1998	R	169 min	Drama, Wa	8.6	Following t	91	Steven Spi	Tom Hanks	Matt Damo	Tom Sizem	Edward Bu	1235804	########
27	https://m.r	The Green	1999	A	189 min	Crime, Dra	8.6	The lives of	61	Frank Dara	Tom Hanks	Michael Cl	David Mors	Bonnie Hu	1147794	########
28	https://m.r	La vita Ã¨ b	1997	U	116 min	Comedy, D	8.6	When an o	59	Roberto Be	Roberto Be	Nicoletta B	Giorgio Ca	Giustino D	623629	########
29	https://m.r	Se7en	1995	A	127 min	Crime, Dra	8.6	Two detect	65	David Finc	Morgan Fre	Brad Pitt	Kevin Spac	Andrew Ke	1445096	########
30	https://m.r	The Silenc	1991	A	118 min	Crime, Dra	8.6	A young F.B	85	Jonathan D	Jodie Foste	Anthony Ho	Lawrence A	Kasi Lemm	1270197	########

https://www.kaggle.com/datasets/harshitshankhdhar/imdb-dataset-of-top-1000-movies-and-tv-shows

DASHBOARD



Year 2015

Genre Drama, War

Stars Emmanuel Affadzi, Abraham Attah Ricky Adelayitor, Andr..

Duration In Minutes 170 min

Title: Beasts of No Nation

META Score 79

IMDB Rating 7.7

No. Of Votes 73,964

Gross $83,861

Description
A drama based on the experiences of Agu, a child soldier fighting in the civil war of an unnamed African country.

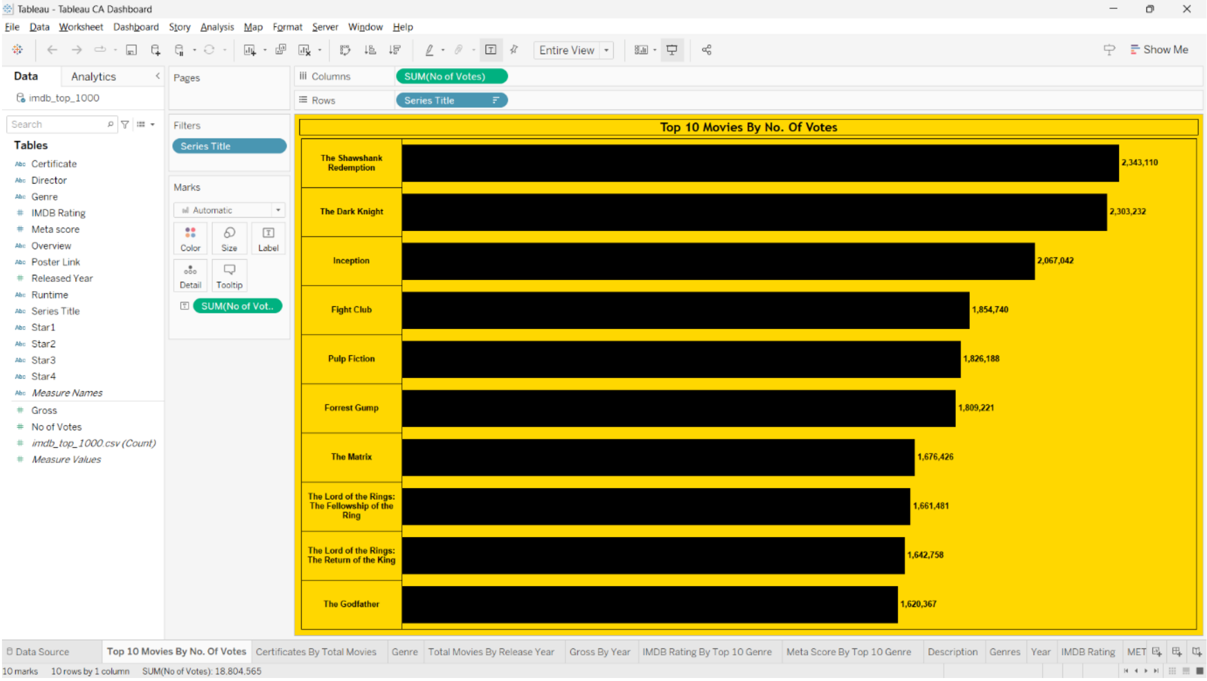
Top 10 Movies By No. Of Votes

The Shawshank Rede..	2,343,110
The Dark Knight	2,303,232
Inception	2,067,042
Fight Club	1,854,740
Pulp Fiction	1,826,188
Forrest Gump	1,809,221
The Matrix	1,676,426
The Lord of the Rings:..	1,661,481
The Lord of the Rings:..	1,642,758
The Godfather	1,620,367

Certificates By Total Movies

U	234
A	197
UA	175
R	146
PG-13	43
PG	37
Passed	34
G	12
Approved	11
TV-PG	3
GP	2
Unrated	1
U/A	1
TV-MA	1
TV-14	1
16	1

Top 10 Movies By No. of Votes



Introduction: This visualization showcases the ten most popular films within the IMDb dataset, based on the number of votes they have received from users. Its purpose is to highlight not only well-rated movies, but those with significant audience engagement on the IMDb platform.

General Description: The visualization is a horizontal bar chart. Each bar represents an individual movie title, with the length of the bar corresponding to the 'No. of Votes' that movie has received. Bars are arranged in descending order, placing the movie with the highest vote count at the top.

Specific Requirements, functions, and formulas: The data was likely sorted by 'No. of Votes' in descending order. For display purposes, the display is probably limited to the top 10 results.

Analysis Results: This visualization reveals which movies have resonated most strongly with IMDb users, leading them to actively cast their votes. Key insights would include:

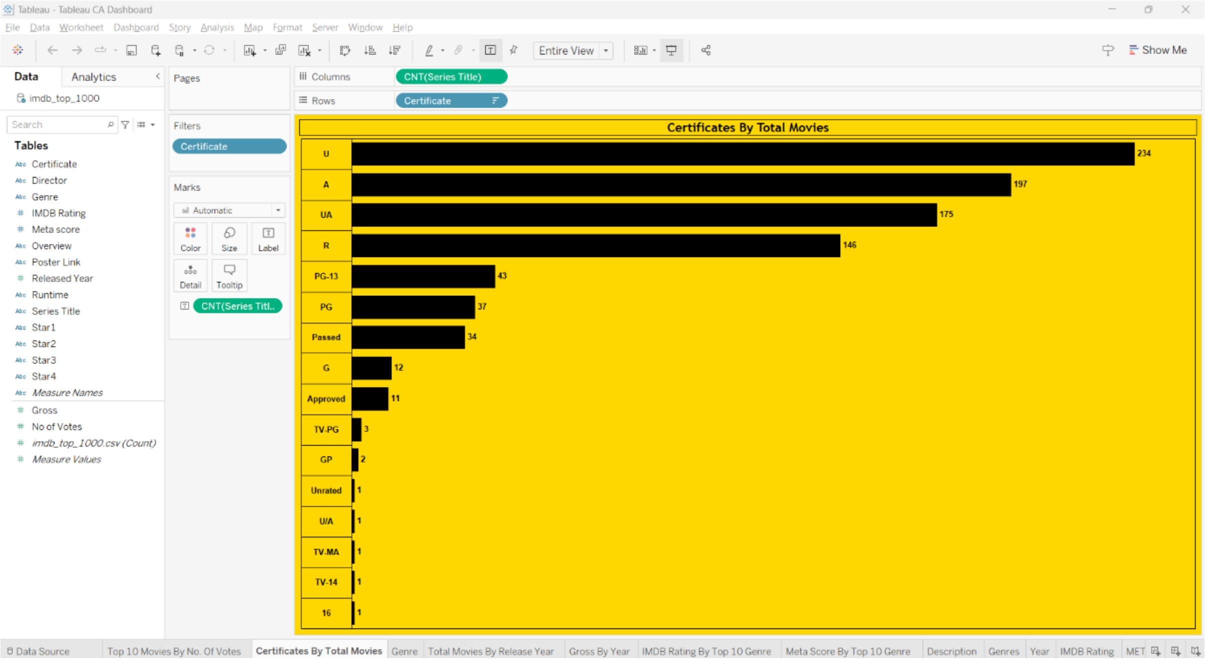
-**Dominant Genres**: Are the most popular movies concentrated within a few genres, or is there diversity?

-**Classic vs. Modern**: Is the list dominated by older releases, newer films, or a mix? This could suggest audience preferences for nostalgia or contemporary storytelling.

-**Critical Acclaim vs. Popularity**: Do the highest-vote films also appear in "best films of all time" lists, or are these crowd favourites distinct from critics' choices?

Visualization: A horizontal bar chart is very effective for this analysis. It visually emphasizes the differences in vote counts between movies, and the ranking format creates an immediate sense of which films stand out as the most popular.

Certificates by Total Movies



Introduction: This visualization explores the distribution of movie certificates (such as G, PG, PG-13, R) within the IMDb dataset. It aims to reveal the relative prevalence of films created for different target audiences, as indicated by their assigned certificate.

General Description: The visualization is a vertical bar chart. Each bar represents a distinct certificate type, and the height of the bar corresponds to the total number of movies within the dataset possessing that certificate.

Specific Requirements, functions, and formulas: The dataset was likely grouped by the 'Certificate' column. A simple count of the occurrences of each certificate type provided the data used to generate the bar heights.

Analysis Results: This visualization sheds light on the types of content that dominate the IMDb dataset. Key findings might include:

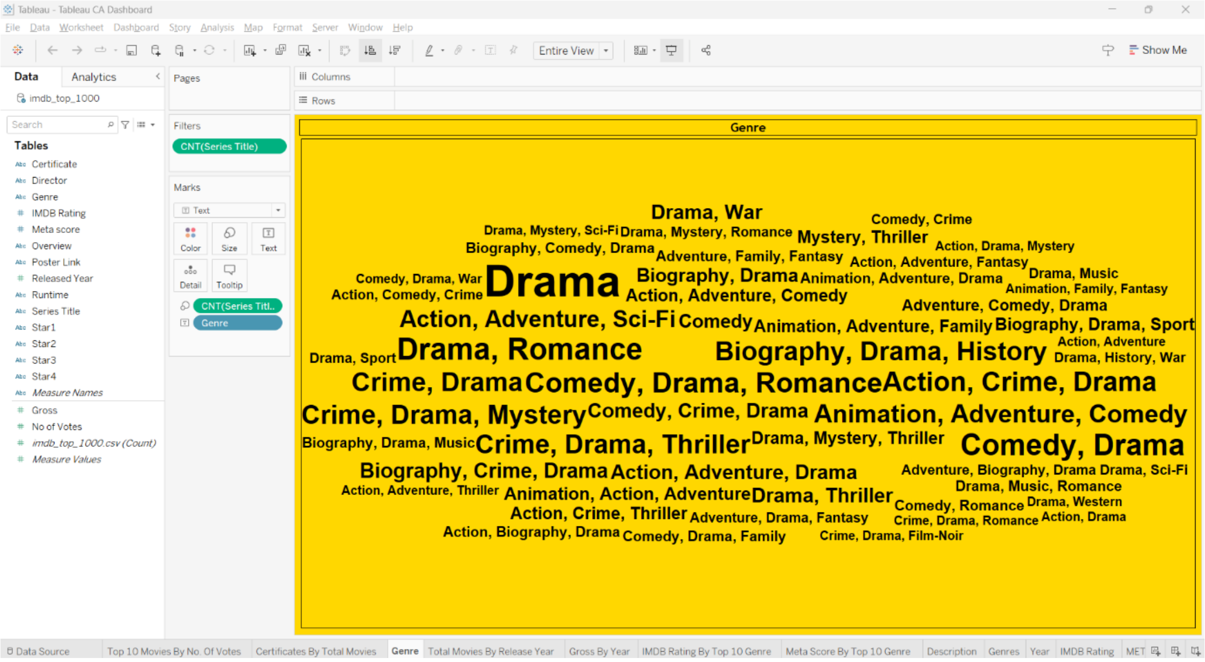
-**Most Common Certificate**: This is the certificate assigned to the largest number of films in the dataset. It suggests the primary target audience the majority of the films cater to.

-**Rarest Certificate**: This identifies a certificate type with significantly fewer associated films. It could point to a niche audience category or films less likely to be included in this specific dataset.

-**Unexpected Distributions**: Are there surprisingly high or low counts for certain certificates? This could indicate a bias in the dataset's composition or a shift in film production trends over time.

Visualization: A vertical bar chart is a clear and effective choice for comparing the frequencies of different categories.

Genre



Introduction: This visualization provides a high-level overview of the different movie genres represented within the IMDb dataset. It aims to highlight the most prominent genres, as well as potentially reveal less common, niche categories.

General Description: The visualization is a word cloud. Each word represents a distinct genre, and the word's font size corresponds to the number of movies within the dataset tagged with that genre.

Specific Requirements, functions, and formulas: The dataset's 'Genre' column was likely parsed to identify unique genres. The frequency of each genre was then calculated, which determines the relative font sizes within the word cloud.

Analysis Results: This visualization quickly reveals the following insights about genre distribution within the dataset:

-**Dominant Genres**: The largest words in the cloud indicate the genres that appear most frequently. This suggests these genres have a strong presence within the dataset.

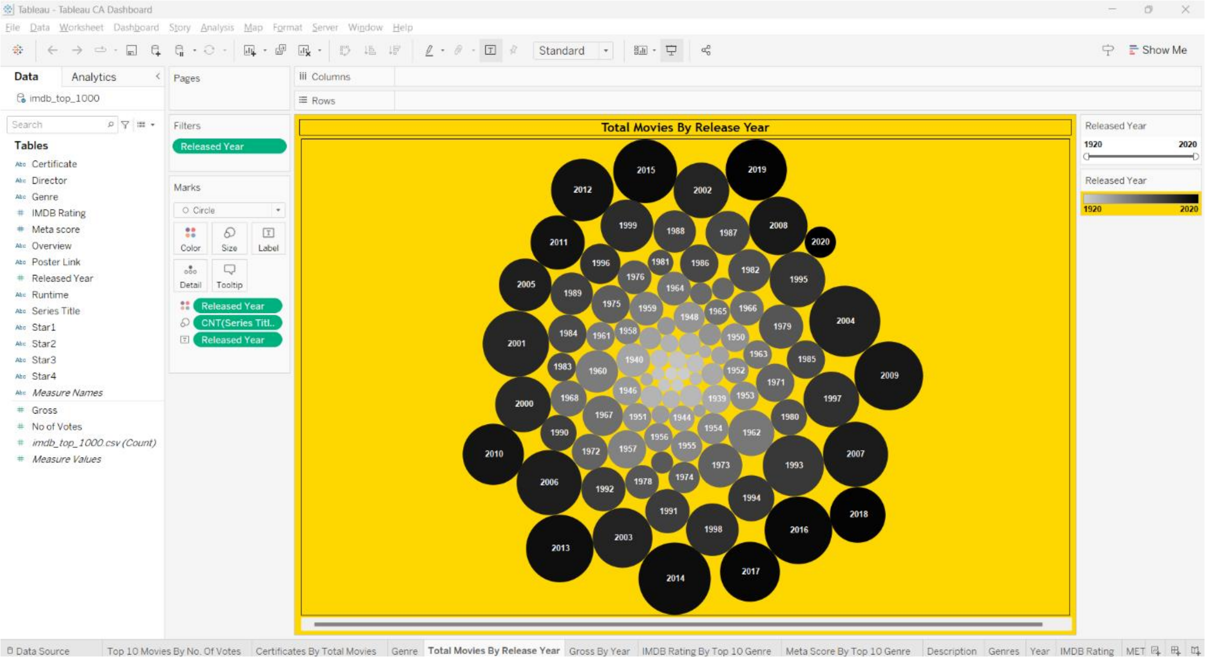
-**Niche Genres:** Smaller words in the cloud may represent less common, more specialized genres

-**Potential Biases:** If certain genres are surprisingly underrepresented, it might indicate that the dataset is not fully comprehensive or has a focus on specific types of films.

Visualization: A word cloud is visually appealing and immediately draws attention to the most common genres. However, it has some limitations for precise analysis:

-**Missing Context**: The word cloud shows how often a genre occurs, but not what kind of films fall within it. A highly popular genre could contain both critically acclaimed and poorly received films.

Total Movies By Release Year



Introduction: This bubble chart provides a multi-dimensional view of the relationship between three variables within your IMDb dataset. It aims to reveal potential correlations, identify outliers, and uncover patterns that might not be apparent using simpler chart types.

General Description: In a bubble chart, each individual film within the dataset is represented by a bubble. The following aspects visually convey information about each film:

Bubble Size: Represents the value of a third variable, which might be duration, number of votes, or any other data point from your dataset.

Specific Requirements, functions, and formulas:

-**Variables**: Identify the three specific data columns from your dataset that are used to determine a bubble's positioning and size.

-**Scaling**: Understand how the bubble chart software maps the data values of each variable to the visual dimensions (position and size) of the chart.

Analysis Results: Carefully analyze the bubble chart, looking for the following insights:

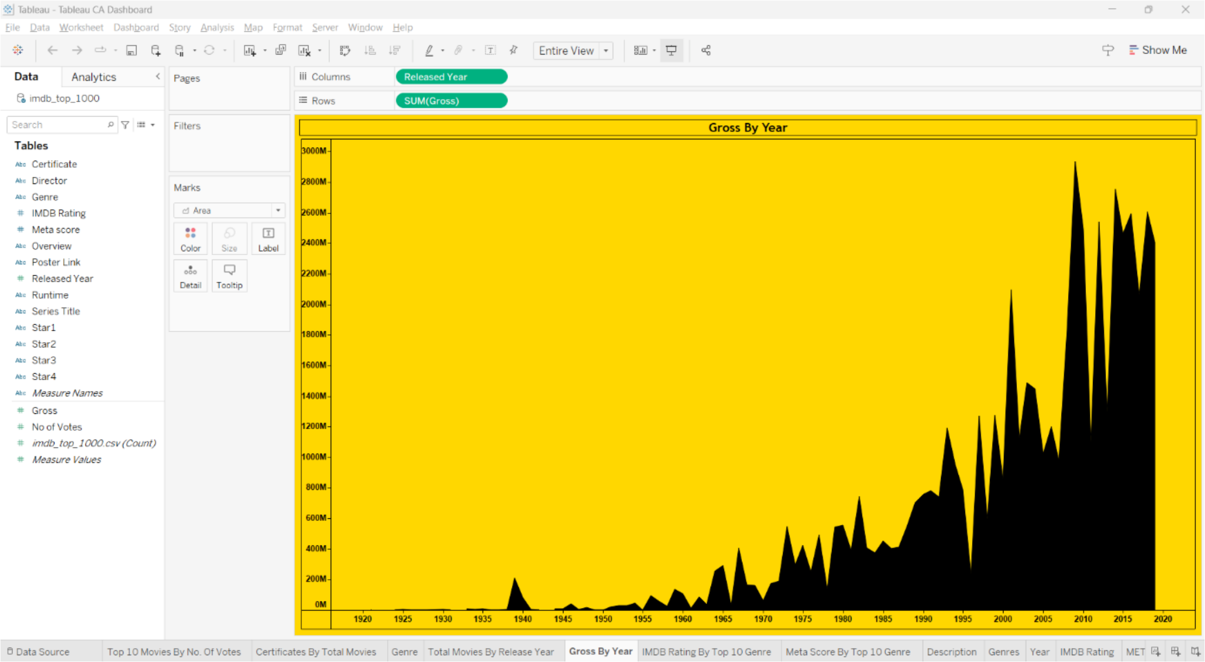
-**Correlations**: Are there noticeable trends? For example, do films with higher IMDB ratings also tend to have higher gross revenue (bubbles may form an upward sloping pattern).

-**Clusters:** Are groups of bubbles concentrated in specific areas of the chart, indicating films that share common traits?

-**Outliers:** Identify unusually large or small bubbles, or bubbles positioned far away from the majority. These might represent exceptional or anomalous films within the dataset.

Visualization: A bubble chart is a powerful tool for comparing multiple variables at once, especially when a large number of data points are involved.

Gross By Year



Columns: Released Year

Rows: SUM(Gross)

Introduction: This area chart depicts the number of films released each year within the IMDb dataset. The visualization aims to reveal trends in film production output over time.

General Description: The chart is a stacked area chart, where each area section represents a distinct category. In this case, there is only one category, so the entire area essentially represents the total number of films released per year. The X-axis displays the release year, and the Y-axis indicates the number of films.

Specific Requirements, functions, and formulas: The data was likely grouped by 'Released Year'. The total number of films released in each year was then calculated and plotted on the Y-axis.

Analysis Results: By observing the areas across different years, we can gain insights into historical trends of film production:

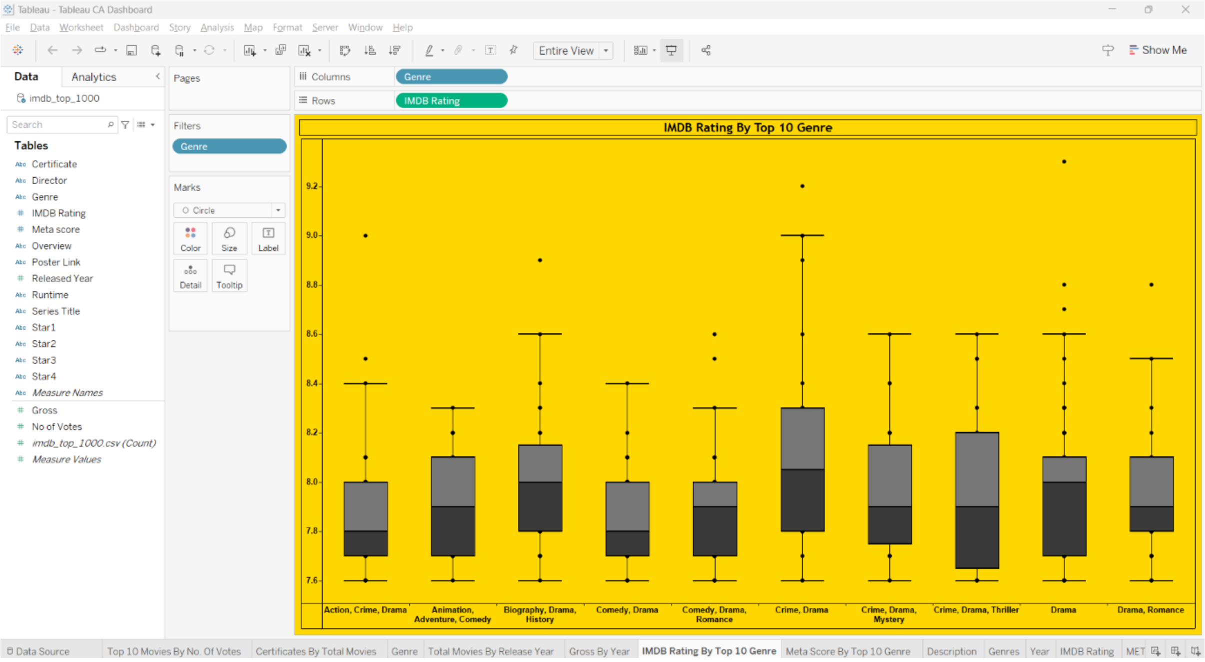
-High Points: Years with larger areas correspond to periods with a higher number of films released. These peaks might be due to various factors, such as economic prosperity, technological advancements in filmmaking, or significant historical events influencing the industry.

-Low Points: Conversely, dips in the area indicate years with lower film output. Potential reasons could include economic downturns, wars, or shifts in audience preferences.

-Long-Term Trends: Is there a general increase or decrease in film production over time? This could be reflected in a gradual upward or downward slope in the overall area across the years.

Visualization: An area chart is a well-suited choice for visualizing trends over time.

IMDB Rating By Top 10 Genre



Introduction: This visualization analyses the distribution of IMDB ratings across the top 10 most frequent genres within your dataset. It uses a box plot format to reveal central tendencies, variability, and potential outliers in the ratings for each genre.

General Description: The chart is a box plot with multiple horizontal boxes, each representing a genre. The boxes themselves are divided by a vertical line in the middle, indicating the median rating for that genre (the middle 50% of ratings fall below or above this line). The horizontal lines extending from the boxes, called whiskers, represent the upper and lower quartiles of the ratings data. Outliers, data points that fall outside a specific range above or below the whiskers, are plotted as individual circles.

Analysis Results: By examining the position and dimensions of the boxes and whiskers, you can learn a lot about how IMDB ratings are distributed within each genre:

-**Median Rating:** The vertical line in the middle of each box represents the median IMDB rating for that genre. Genres with higher medians tend to receive more favorable user ratings overall.

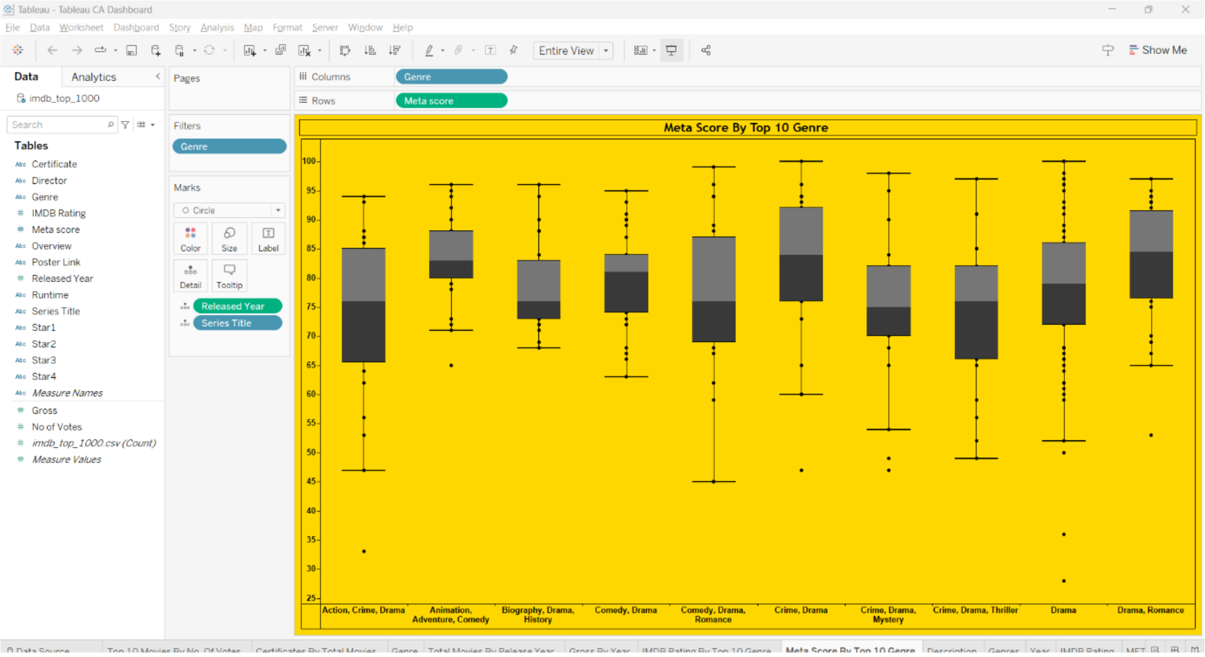
-**Spread of Ratings:** The width of the box indicates the interquartile range (IQR), which represents the middle 50% of the ratings data. A wider box suggests a larger spread of ratings within that genre, while a narrow box indicates more consistent scores.

-**Whiskers:** The whiskers extend from the box to Q1 (lower quartile) and Q3 (upper quartile). The length of the whiskers shows how far the remaining 25% of ratings fall above and below the middle 50%. Longer whiskers suggest a wider distribution of ratings within a genre.

-**Outliers:** The data points plotted as individual circles are outliers, values that fall outside the range of 1.5 IQR below Q1 or above Q3. These could represent films that received exceptionally high or low ratings compared to others within their genre.

Visualization: A box plot is a concise and informative way to compare distributions of data across multiple categories.

Meta Score By Top 10 Genre



Introduction: This visualization explores the distribution of Meta scores within the top 10 most frequent genres in your dataset. It utilizes a box plot format to reveal medians, ranges, and potential outliers for Meta scores across these genres.

General Description: The chart is a box plot with multiple horizontal boxes, each representing a genre. The boxes are divided by a vertical line in the middle, indicating the median Meta score for that genre (half of the scores fall below or above this line). The horizontal lines extending from the boxes, called whiskers, represent the upper and lower quartiles of the Meta score data. Data points outside a specific range above or below the whiskers are plotted as individual circles, signifying outliers.

-Analysis Results: By examining the position and dimensions of the boxes and whiskers, you can learn a lot about how Meta scores are distributed within each genre:

-Median Score: The vertical line in the middle of each box represents the median Meta score for that genre. Genres with higher medians tend to receive more favorable scores from professional critics.

-Spread of Scores: The width of the box indicates the interquartile range (IQR), which represents the middle 50% of the Meta score data. A wider box suggests a larger spread of scores within that genre, while a narrow box indicates more consistent scores.

-Whiskers: The whiskers extend from the box to Q1 (lower quartile) and Q3 (upper quartile). The length of the whiskers shows how far the remaining 25% of ratings fall above and below the middle 50%. Longer whiskers suggest a wider distribution of scores within a genre.

-Outliers: The data points plotted as individual circles are outliers, values that fall outside the range of 1.5 IQR below Q1 or above Q3. These could represent films that received exceptionally high or low scores from critics compared to others within their genre.

Visualization: A box plot is a concise and informative way to compare distributions of data across multiple categories.

Analysis List and Outcomes

Analysis 1: Top 10 Movies by Number of Votes (Bar Chart)

Outcome: Classic films like "The Shawshank Redemption" and "The Dark Knight" topped the list, along with more recent hits like "Inception" and "Fight Club". This suggests enduring popularity for critically acclaimed titles, alongside blockbusters that resonated with a wide audience.

Analysis 2: Certificates By Total Movies (Bar Chart)

Outcome: The most frequent movie certificates were "R" and "PG-13", indicating a dominance of mature content, likely targeting adult audiences.

Analysis 3: Genre (Word Cloud)

Outcome: The word cloud visualization reinforces the dominance of Drama, Comedy, and Action genres within the dataset. The font sizes indicate that these are the most frequent genres.

Analysis 4: Meta Score By Top 10 Genre (Box Plot)

Outcome: This box plot reveals that Drama, Comedy, and Animation genres generally received higher median Meta scores, indicating critical favor. Action and Crime tended towards lower median scores. The spread (width) of the boxes suggests that ratings for some genres, like Comedy and Drama, were more variable than others, such as Animation.

Analysis 5: IMDB Rating by Top 10 Genre (Box Plot)

Outcome: Similar to the Meta score analysis, this box plot suggests that Drama, Comedy, and Animation genres tended towards higher median user ratings on IMDb. There's a wider spread in ratings for some genres (Comedy, Drama) compared to others (Animation), indicating more diverse user opinions within those categories.

Analysis 6: Gross By Year (Area Chart)

Outcome: The area chart suggests an overall increase in film production over time. There might be peaks and valleys in film releases, potentially corresponding to historical trends or economic cycles in the film industry.

<u>THANK YOU</u>