

Data Visualization

(Final assignment)

Bita Naserfarahmand

Student ID: 2021380075

Email: bita.nf@gmail.com

Data Visualization class

Content:

<i>1. Introduction.....</i>	<i>3</i>
<i>2. Goal.....</i>	<i>4</i>
<i>3. Data Preprocessing.....</i>	<i>6</i>
<i>4. Visualizations.....</i>	<i>7</i>
<i>5. Findings & Conclusion.....</i>	<i>13</i>
<i>6. References.....</i>	<i>16</i>

1) Introduction:

Questions:

- 1. What is your data?*
- 2. what is your question and what do you want to achieve by visualizing the data?*
- 3. How will you visualize your data?*

Apparatus:

Anaconda 3 , Jupyter notebook

2)Goal:

1. What is your data?

The dataset, IMDB Dataset of Top 1000 Movies and TV Shows, provides detailed information about the top-rated movies and TV shows on IMDB. This dataset includes crucial attributes that allow for in-depth analysis of trends in movie genres, ratings, popularity, and other factors. It is sourced from IMDB and is suitable for projects involving data visualization, exploratory data analysis (EDA), and machine learning.

Key Features of the Dataset:

1. Data Size:

Contains data for 1,000 top-rated movies and TV shows.

2. Attributes/Columns:

Poster_Link: URL to the poster image of the movie/TV show.

Series_Title: Title of the movie or TV show.

Released_Year: Year the movie or TV show was released.

Certificate: Certification of the movie/TV show (e.g., PG, R).

Runtime: Duration of the movie/TV show in minutes.

Genre: Genre(s) of the movie/TV show (e.g., Drama, Action, Comedy).

IMDB_Rating: IMDB rating of the movie/TV show (out of 10).

Overview: A brief description or synopsis.

Meta_score: MetaCritic score for the movie/TV show.

Director: Name(s) of the director(s).

Star1, Star2, Star3, Star4: Names of the main actors/actresses.

No_of_Votes: Total number of votes received on IMDB.

Gross: Worldwide gross earnings (in USD, where available).

	Poster_Link	Series_Title	Released	Certificate	Runtime	Genre	IMDB_Rat	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Votes	Gross
1		Two imprints	1994 A	142 min	Drama	9.3	Two imprints	80	Frank Darab	Tim Robbins	Morgan Freeman	Bob Odenkirk	William B. Davis	2343110	#####	
2	https://m...	The Godfather	1972 A	175 min	Crime, Drama	9.2	An organized crime	100	Francis Ford Coppola	Marlon Brando	Al Pacino	James Caan	Diane Keaton	1620367	#####	
3	https://m...	The Dark Knight	2008 UA	152 min	Action, Crime	9	When the	84	Christopher Nolan	Christian Bale	Heath Ledger	Aaron Eckhart	Michael Caine	2303232	#####	
4	https://m...	The Godfather Part II	1974 A	202 min	Crime, Drama	9	The early l	90	Francis Ford Coppola	Al Pacino	Robert De Niro	Diane Keaton		1129952	#####	
5	https://m...	12 Angry Men	1957 U	96 min	Crime, Drama	9	A jury hold	96	Sidney Lumet	Henry Fonda	Lee Remick	John Fiedler		689945	4,360,000	
6	https://m...	The Lord of the Rings: The Fellowship of the Ring	2003 U	201 min	Action, Adventure	8.9	Gandalf ar	94	Peter Jackson	Elijah Wood	Viggo Mortensen	Orlando Bloom		1542758	#####	
7	https://m...	Pulp Fiction	1994 A	154 min	Crime, Drama	8.9	The lives o	94	Quentin Tarantino	John Travolta	Uma Thurman	Samuel L. Jackson	Bruce Willis	1826188	#####	
8	https://m...	Schindler's List	1993 A	195 min	Biography, Drama	8.9	In German	94	Steven Spielberg	Liam Neeson	Ralph Fiennes	Ben Kingsley	Caroline Goodall	1213505	#####	
9	https://m...	Inception	2010 UA	148 min	Action, Adventure	8.8	A thief wh	74	Christopher Nolan	Leonardo DiCaprio	Joseph Gordon-Levitt	Elliot Page	Ken Watanabe	2067042	#####	
10	https://m...	The Fight Club	1999 A	139 min	Drama	8.8	An insomn	66	David Fincher	Brad Pitt	Edward Norton	Meat Loaf	Zach Grenier	1854740	#####	
11	https://m...	The Lord of the Rings: The Two Towers	2002 U	178 min	Action, Adventure	8.8	A meek h	92	Peter Jackson	Elijah Wood	Ian McKellen	Orlando Bloom	Sean Bean	1661481	#####	
12	https://m...	The Forrest Gump	1994 UA	142 min	Drama, Romance	8.8	The presid	82	Robert Zemeckis	Tom Hanks	Robin Williams	Wendie Malick	Sally Field	1809221	#####	
13	https://m...	Il buono, il brutto, il cattivo	1966 A	161 min	Western	8.8	A bounty h	90	Sergio Leone	Clint Eastwood	Frank John Hughes	Lee Van Cleef	Aldo Giuffrè	688390	6,100,000	
14	https://m...	The Lord of the Rings: The Return of the King	2003 UA	179 min	Action, Adventure	8.7	While Frodo	87	Peter Jackson	Elijah Wood	Ian McKellen	Viggo Mortensen	Orlando Bloom	1485555	#####	
15	https://m...	The Matrix	1999 A	136 min	Action, Sci-Fi	8.7	When a be	73	Lana Wachowski	Lilly Wachowski	Keanu Reeves	Laurence Fishburne	Carrie-Anne Moss	1876426	#####	
16	https://m...	Goodfellas	1990 A	146 min	Biography, Crime, Drama	8.7	The story i	90	Martin Scorsese	Robert De Niro	Liam Neeson	Joe Pesci	Lorraine Bracco	1020727	#####	
17	https://m...	Star Wars: The Force Awakens	1980 UA	124 min	Action, Adventure	8.7	After the F	82	J.J. Abrams	Mark Hamill	Harrison Ford	Carrie Fisher	Billy Dee Williams	1159315	#####	
18	https://m...	One Flew Over the Cuckoo's Nest	1975 A	133 min	Drama	8.7	A criminal	83	Milos Forman	Jack Nicholson	Florence LaSeue	Michael Berryman	Peter Brocco	918088	#####	
19	https://m...	Hamilton	2020 PG-13	160 min	Biography, Musical	8.6	The real lif	90	Thomas Kail	Lin-Manuel Miranda	Phillipa Soo	Leslie Odom Jr.	Renée Elise Goldsberry	55291	#####	
20	https://m...	Gisaengchung	2019 A	132 min	Comedy, Crime, Drama	8.6	Greed and	96	Bong Joon-ho	Kang-ho Song	Lee Sun-kyo	Choi Woo-young	Choi Hae-ri	552778	#####	
21	https://m...	Soorai P	2020 U	153 min	Drama	8.6	Nedumaaran	86	Nedumaaran	Rajaguru	Sudha Konar	Suriya	Madhavan	54995	#####	
22	https://m...	Interstellar	2014 UA	169 min	Adventure, Drama, Sci-Fi	8.6	A team of	74	Christopher Nolan	Matthew McConaughey	Anne Hathaway	Jessica Chastain	Mackenzie Davis	1512360	#####	
23	https://m...	Cidade de Deus	2002 A	130 min	Crime, Drama	8.6	In the slum	79	Fernando Meirelles	Katia Lin	Alexandre Rodrigues	Leandro Firmino	Mathias Resende	699256	7,563,397	
24	https://m...	Sen to Chihiro no Kamikakushi	2001 U	125 min	Animation	8.6	During her	96	Hayao Miyazaki	Miyu Irino	Rumi Hiiragi			651376	#####	
25	https://m...	Saving Private Ryan	1998 R	169 min	Drama, War	8.6	Following i	91	Steven Spielberg	Tom Hanks	Matthew Damon	Tom Sizemore	Edward Burns	1235804	#####	
26	https://m...	The Green Mile	1999 A	189 min	Crime, Drama	8.6	The lives o	61	Frank Darab	Tom Hanks	Michael Clarke Duncan	David Morse	Bonnie Hunt	1147794	#####	
27	https://m...	La vita è bella	1997 U	116 min	Comedy, Crime, Drama	8.6	When an c	59	Roberto Benigni	Roberto Benigni	Nicoletta Marzi	Giorgio Capecchi	Giustino D'Amico	623629	#####	
28	https://m...	Se7en	1995 A	127 min	Crime, Drama	8.6	Two detec	65	David Fincher	Morgan Freeman	Brad Pitt	Kevin Spacey	Andrew Keels	1445096	#####	

2. what is your question and what do you want to achieve by visualizing the data?

My goal in this project is to investigate how the popularity of movie genres has evolved over time, using data from the IMDB Top 1000 Movies dataset. By analyzing trends in audience preferences—such as the number of votes and ratings across different genres in various time periods—I aim to uncover significant shifts in popularity. Through this analysis, I will also attempt to understand the underlying factors that may have contributed to these changes, such as cultural influences, advancements in filmmaking technology, or changes in audience demographics. The insights gained from this exploration can provide a deeper understanding of how cinema has evolved to meet audience expectations over the decades.

3. How will you visualize your data?

To analyze the changes in genre popularity over time and identify the key contributors, I use the following visualizations:

1. Stacked Area Chart:

- This chart shows the overall popularity of all genres over time based on the number of votes. It provides a clear view of how audience preferences evolved collectively.

2. Stacked Area Chart for Top 5 Genres:

- A focused version of the stacked area chart that highlights only the top 5 genres, offering a more detailed view of the most popular genres over time.

3. Line Chart for Top 5 Genres:

- A line chart for the top 5 genres, allowing precise comparisons of their trends without stacking.

4. Heatmap:

- A heatmap that visualizes the popularity of genres across years. The intensity of color represents the level of popularity, making it easy to identify peaks and declines.

5. Small Multiples (Individual Line Charts):

- Separate line charts for the top 6 genres, displayed side-by-side. This avoids overlapping and provides clarity for each genre's trend.

6. Interactive Plot (Plotly):

- An interactive line plot created with Plotly, allowing users to dynamically explore genre popularity trends over time.

7. Drill-Down Analysis:

- Identifying the top years for each genre and drilling down to highlight the specific movies that contributed to popularity spikes.

8. Bar Chart for Top Movies in Key Years:

- A bar chart showcasing the top 10 movies driving genre popularity in significant years. This highlights individual contributions to overall trends.

3) Data Preprocessing:

To prepare the dataset for analysis and ensure the accuracy of the results, I performed several data preprocessing steps:

1. Handling Missing Values:

- Rows with missing or incomplete data in critical columns, such as Released_Year, Genre, and No_of_Votes, were removed to avoid inaccuracies during analysis.

2. Converting Data Types:

- The Released_Year column, which initially contained non-numeric or inconsistent values, was cleaned and converted into integers to facilitate chronological analysis.

3. Splitting Genres:

- Movies with multiple genres listed in a single row were split into separate rows for each genre. This allowed for a more granular analysis of genre popularity trends.

4. **Standardizing Columns:**

- Text columns, such as Genre, were processed to ensure consistency by handling formatting issues like extra spaces or special characters.

5. **Aggregating Data:**

- The cleaned data was grouped by Released_Year and Genre to calculate the total number of votes for each combination. This aggregation helped in identifying genre trends over time.

6. **Cleaning Monetary Values** (if applicable):

- Columns like Gross, which contained monetary values with formatting issues (e.g., commas), were standardized and converted into numeric data types for analysis.

These preprocessing steps ensured that the dataset was clean, consistent, and suitable for generating accurate insights and meaningful visualizations.

4) Visualization:

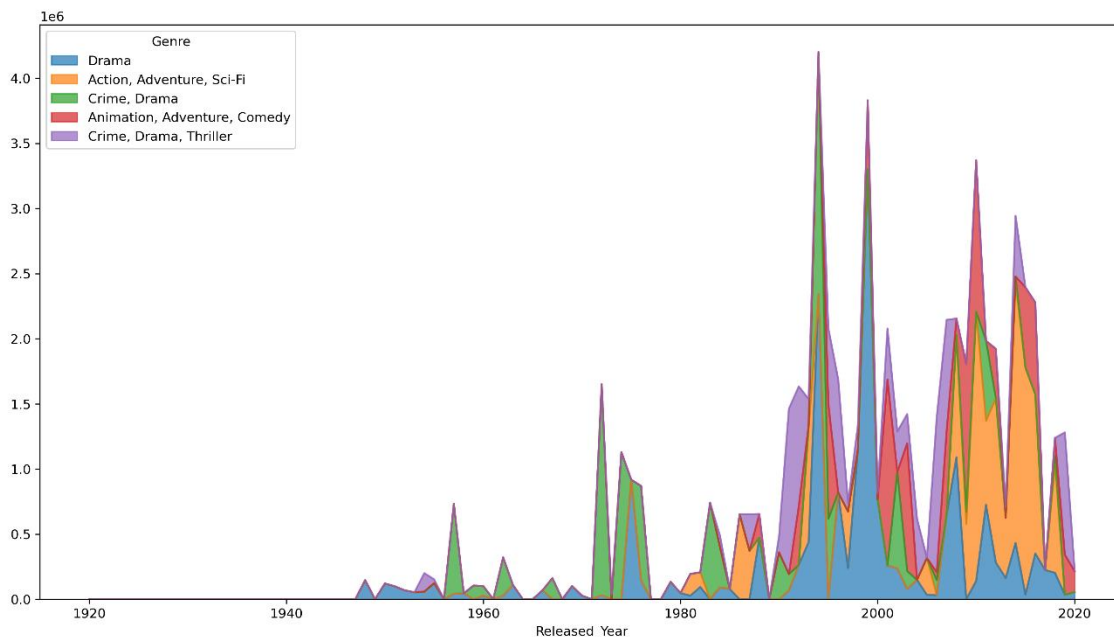
1. **Stacked Area Chart: Total Votes by Genre Over Time**

Description:

Initially, I visualized all genres, but the chart was too cluttered. I narrowed it down to the top 5 genres based on votes. This chart shows the cumulative votes for these genres over time, with each genre represented by a distinct color.

Analysis:

- Drama has been consistently popular across decades.
- Action and Sci-Fi saw significant growth in the late 20th century, peaking in the 2000s.
- Other genres like Animation and Crime show periodic spikes rather than sustained popularity.



2. Individual Line Charts: Top Genres Over Time

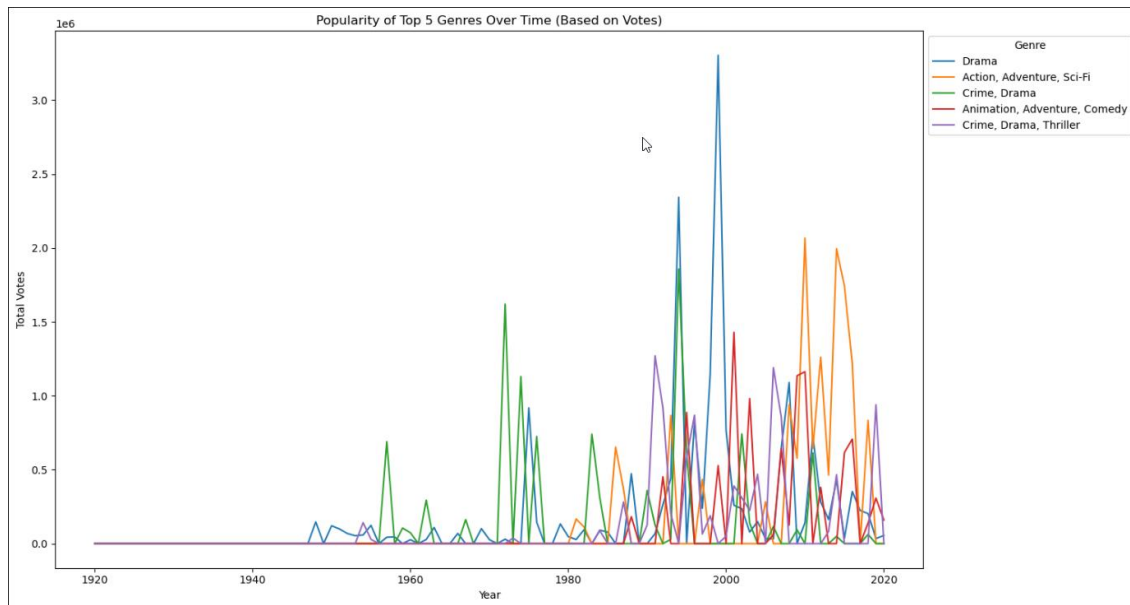
Description:

This line plot shows the popularity of the top five genres over time, measured by the total votes they received. Unlike the stacked area chart, this visualization separates the genres, making it easier to track individual trends.

Analysis:

- Drama consistently dominates across the timeline, with notable peaks around the early 2000s.
- Action, Adventure, Sci-Fi shows significant spikes in the late 1990s and early 2000s, indicating the influence of blockbuster films in that period.

- Crime, Drama has a steady rise with peaks aligning with critical acclaim and audience favorites.
- Animation, Adventure, Comedy saw a notable increase starting from the 1990s, likely tied to the rise of major animation studios.
- Crime, Drama, Thriller shows intermittent spikes, often tied to specific iconic movies.



3. Facet Grid Line Plots

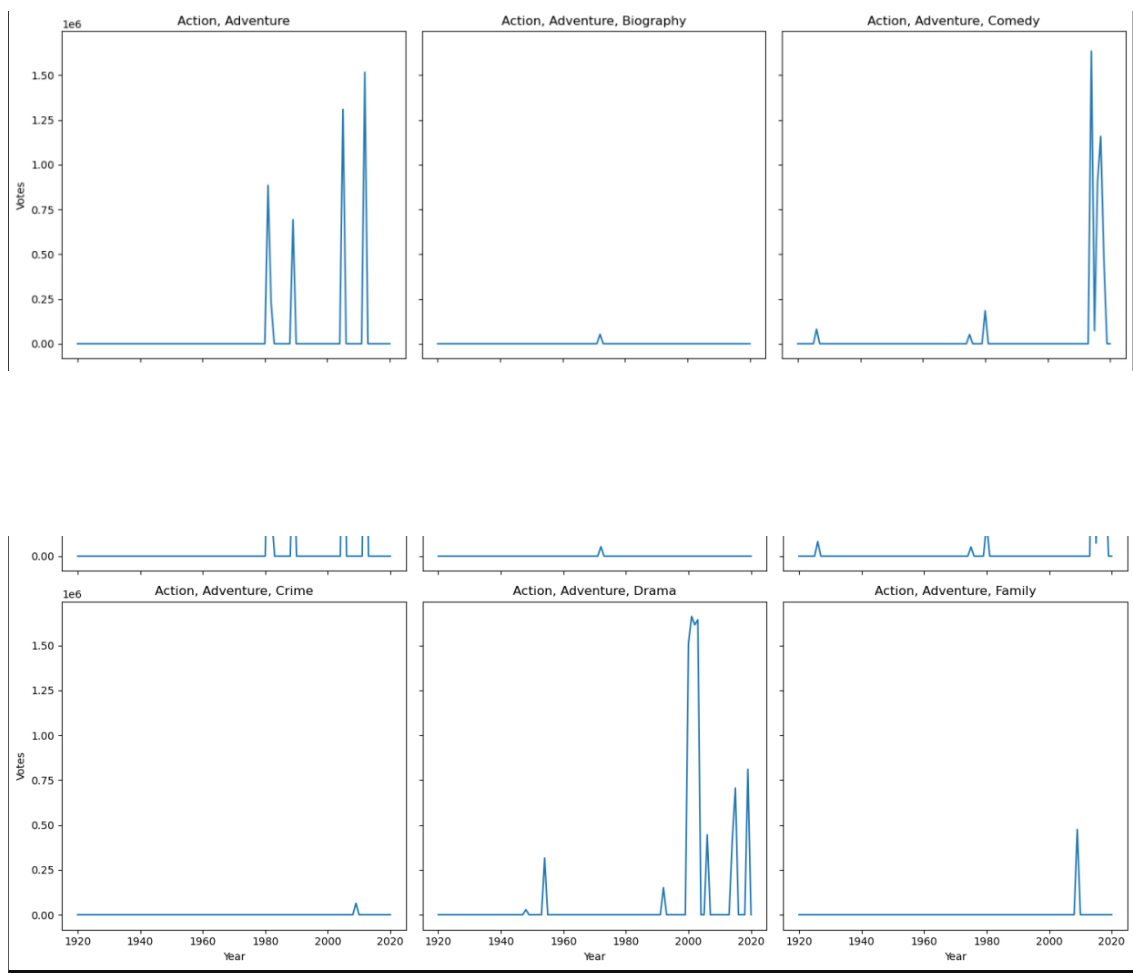
Description:

This visualization consists of separate line plots for different genres, splitting them into manageable sections for easier analysis. Each subplot represents the trend of total votes over time for a specific genre or combination of genres. The years are on the x-axis, and the total votes are on the y-axis.

Analysis:

- Action, Adventure, Drama: Shows a significant rise in popularity starting around the 2000s, peaking in the late 2000s.

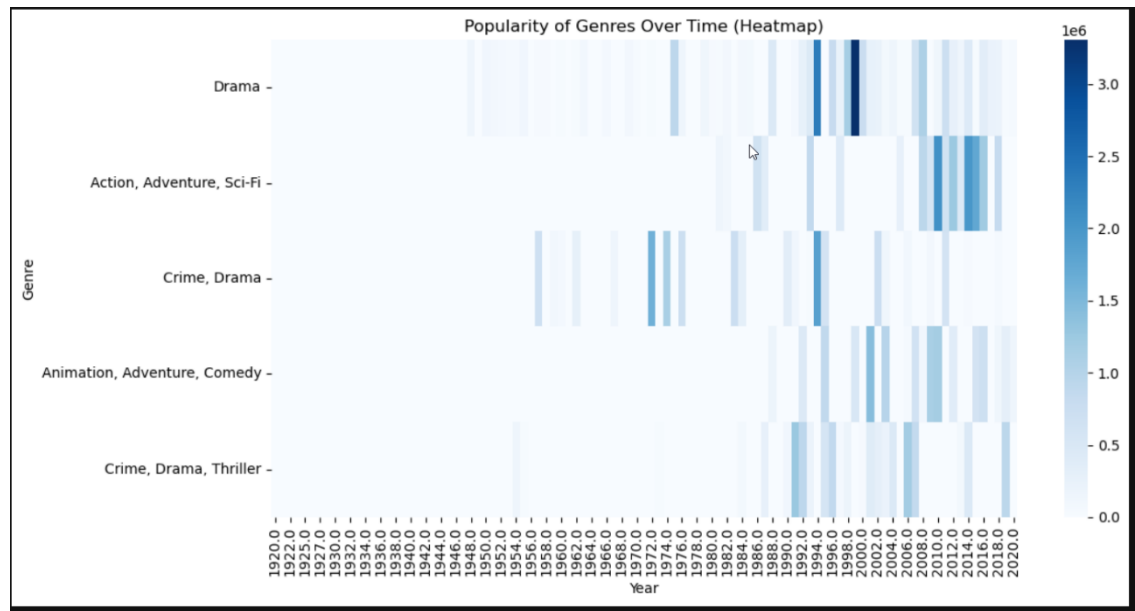
- Action, Adventure, Comedy: Experiences sporadic spikes, with noticeable peaks in recent years.
- Action, Adventure: Demonstrates a consistently increasing trend in the late 1990s and 2000s.
- Action, Adventure, Family: Limited activity but has notable peaks in the early 2000s.
- Action, Adventure, Biography: Shows minimal activity overall, indicating niche popularity.
- Action, Adventure, Crime: Exhibits limited activity, with a few sharp peaks.



4. Heatmap: Genre Popularity Over Time

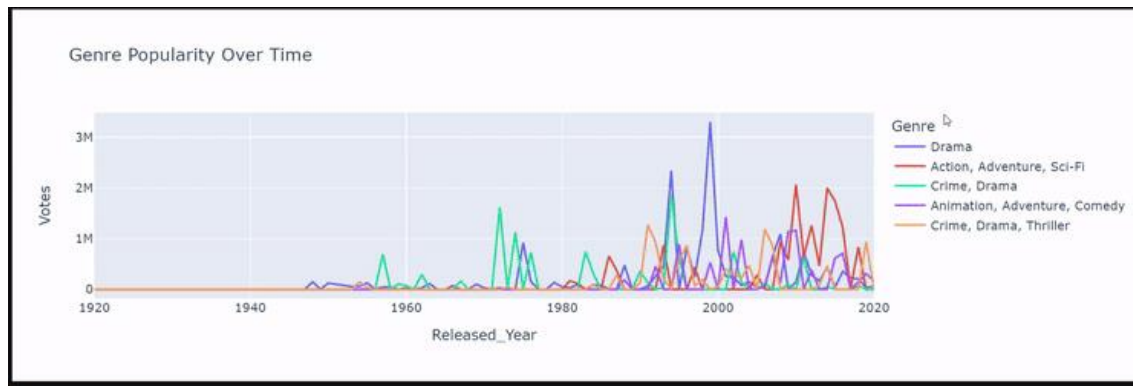
Description: This heatmap visualizes the intensity of popularity (votes) for each genre over the years.

Analysis: The color intensity reveals significant periods for certain genres. For instance, Action genres became notably popular in the 2000s, while Drama maintained steady popularity.



5. Interactive Line Plot: Genre Popularity Over Time

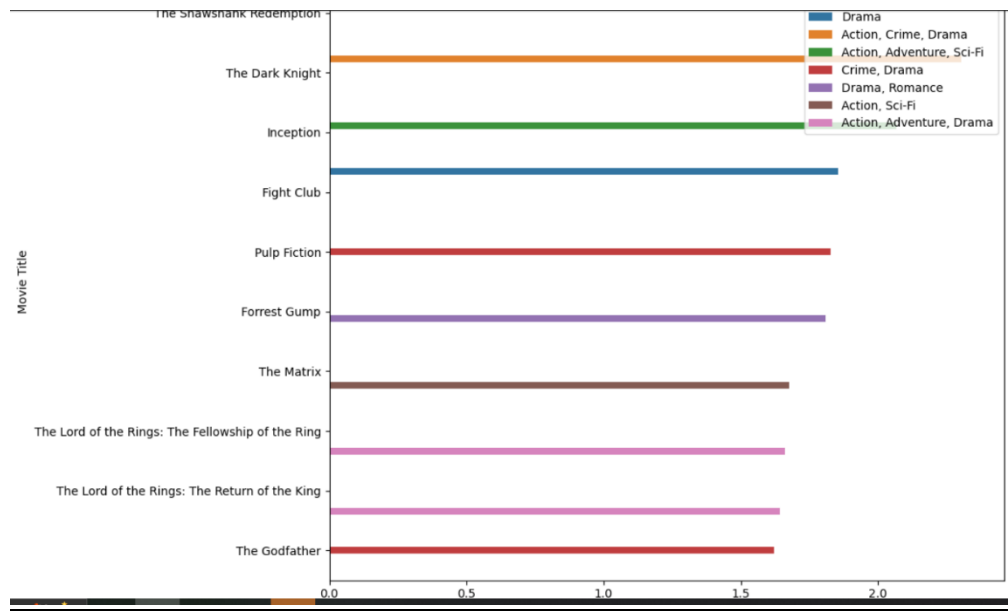
This visualization presents the same analysis as the static line plot but offers an interactive interface. Users can hover over points, filter genres, and explore trends in votes for top genres over the years dynamically. This interactive feature enhances user engagement and makes it easier to analyze specific trends.



6. Drill-Down Analysis: Top Movies in Key Years

Description: This bar chart displays the top movies contributing to genre popularity in key years.

Analysis: Movies like *The Shawshank Redemption* and *The Dark Knight* stand out as major contributors, reflecting how individual blockbusters significantly influence genre trends.



8) Findings & Conclusion

Key Observations:

1. Dominant Genres in Recent Years (Post-2000):

- Drama stands out as a consistently popular genre, peaking in the 2000s.
- Genres like Action, Adventure, Sci-Fi show significant spikes, indicating their rising popularity, likely due to blockbuster franchises and advancements in special effects.

2. Fluctuations in Popularity:

- Genres such as Crime, Drama, Thriller have had intermittent popularity spikes, suggesting they may rely on specific hit movies rather than consistent demand.
- Animation, Adventure, Comedy appears less consistent but shows notable peaks, potentially due to major animated releases.

3. Historical Trends (1920-1980):

- Prior to the 1980s, most genres had low popularity (indicated by fewer votes). This is likely due to limited global accessibility to older movies and lower participation in IMDb ratings for films from earlier decades.

4. Emergence of Multigenre Films:

- Many genres (e.g., Action, Adventure, Sci-Fi) are combinations, reflecting a trend toward blending elements to appeal to broader audiences. This might explain their larger spikes compared to single-genre films.

5. Sharp Peaks:

- The spikes for certain genres likely correspond to iconic or critically acclaimed films that defined or reinvigorated those genres during their release years.

Insights & Hypotheses

- Technology Influence: The rise of Action, Adventure, Sci-Fi after the 1980s aligns with technological advancements in filmmaking (e.g., CGI, large-scale productions).
- Drama's Stability: Drama's steady performance across decades suggests its universal appeal and connection with storytelling.
- Seasonal Trends: Peaks for genres like Animation, Adventure, Comedy could correspond to family-oriented releases around holidays.

After the Drill down visualization

1. Dominant Movies by Genre:

- Drama:
 - *The Shawshank Redemption* leads in votes, reflecting its universal appeal and critical acclaim.

- *Forrest Gump* also represents a significant contribution to the Drama genre's popularity.
- Action/Adventure/Sci-Fi:
 - *The Dark Knight*, *Inception*, and *The Matrix* showcase the dominance of visually striking, big-budget blockbusters.
 - The *Lord of the Rings* trilogy (*The Fellowship of the Ring* and *The Return of the King*) represents another major driver of popularity in this category.
- Crime/Drama:
 - *The Godfather* and *Pulp Fiction* stand out as iconic contributions that elevated this genre's popularity in their respective years.

2. Multigenre Movies:

- Many of these top movies span multiple genres, such as *Action*, *Adventure*, *Sci-Fi* or *Drama*, *Romance*. This trend highlights the versatility and broader appeal of films that cater to diverse audience preferences.

3. Vote Distribution:

- *The Shawshank Redemption* leads by a significant margin, underscoring its enduring popularity over decades.
- Other movies like *The Dark Knight* and *Inception* follow closely, showing how modern films attract larger audiences due to wider global access and engagement on platforms like IMDb.

4. Impact of Franchise Films:

- Franchise movies (*The Lord of the Rings*) have a recurring impact, indicating their role in sustaining popularity for the Adventure genre over multiple years.

Insights and Hypotheses

1. Critical Acclaim and Popularity:
 - Highly-rated movies like *The Shawshank Redemption* and *The Godfather* not only excel in quality but also drive long-term engagement, as seen by their high vote counts.
2. Genre Fusion as a Strategy:
 - Combining genres such as Action, Adventure, and Sci-Fi seems to attract larger audiences, especially in blockbuster movies.
3. Franchise Dominance:
 - Films part of larger franchises (e.g., *The Lord of the Rings*) consistently perform well due to their established fanbase and massive marketing efforts.

9)References:

1. **Kaggle:** The dataset for this project was sourced from Kaggle, specifically from the [IMDB Dataset of Top 1000 Movies and TV Shows](#).
2. **YouTube:** YouTube tutorials were helpful in understanding data visualization techniques and Python libraries like Matplotlib, Seaborn, and Plotly. For instance:
Corey Schafer - Matplotlib Tutorial
StatQuest - Data Visualization
3. **Plotly Documentation:** The official Plotly documentation was used to understand how to create interactive visualizations.
(<https://plotly.com/python/>)

4. **Seaborn Documentation:** For guidance on creating heatmaps and line plots, the Seaborn library documentation was referenced.
(<https://seaborn.pydata.org/>)
5. **Matplotlib Documentation:** For stacked area charts and other basic visualizations, the Matplotlib documentation was used.
(<https://matplotlib.org/stable/contents.html>)