Milestone 1

Project Title: Analyzing the Evolution and Impact of Animated Movies Worldwide

Team: MoscowMickeyMule | **Members:** Lina Bousbina, Michael Burch, Alessandro Salo

---

Dataset

52,000 Animation Movies Dataset:

The following describes the preprocessing steps conducted for a Kaggle dataset on animation movies. The dataset has **51945** entries, with **23 numerical or string-based columns,** such as revenue, runtime, original_title or spoken_languages.

MovieLens Dataset:

The **MovieLens "ml-latest"** dataset, was found on the official GroupLens website. It contains over 33 million 5-star user ratings and 2.3 million free-text tags for approximately 86,000 movies, provided by around 331,000 users. The dataset is organized into several CSV files, including `ratings.csv`, `movies.csv`, `tags.csv`, and `links.csv`, with fields such as movie titles, genres, user ratings, tags, and external identifiers (IMDb, TMDb).

The data is well-structured, complete, and encoded in UTF-8. Missing values are minimal, mainly in the `tags.csv` and `links.csv` files, and have been handled using basic data cleaning methods like `dropna()`. Preprocessing required for this project includes extracting release years from movie titles, filtering movies by genre (Animation), and merging relevant data files. Overall, the dataset is of high quality and requires only light preprocessing, making it ideal for data visualization and educational analysis.

IMBD Dataset:

The IMBD dataset comes in three parts: basic movie details, movie ratings, and logistic details. The dataset can be found on IMBDs developer platform for non-commercial use. This dataset includes TV-shows and movies of all kinds. As a result, preprocessing is required to isolate animated movies.

Animation Studio Dataset:

This Kaggle dataset contains the names, founding year, and country of origin of **348 animation studios**. The dataset is small but complete. As a result, the dataset requires minimal preprocessing. Unfortunately, the dataset only lists studios which uniquely release animations , meaning some big names (i.e. Disney or Dreamworks) are missing.

Problematic

The primary aim of our visualization is to provide users with a **clear understanding of the trends in animation movies and studios, both over time and across countries**. We explore how the production and popularity of animated films have evolved, and how these trends may correlate with significant global events such as the release of breakthrough animation software, the emergence of influential studios, or major award wins. By examining changes in viewer ratings, tagging behavior, and release frequency, we hope to uncover meaningful patterns that reflect both artistic and technological shifts in the industry.

Our motivation lies in animation's **unique position at the crossroads of storytelling, innovation, and global culture.** Whether it's the rise of Japanese anime, the global success of Pixar, or the digital revolution in CGI, animation often mirrors broader shifts in media and technology. This visualization is intended for film enthusiasts, media researchers, and industry professionals seeking to understand how animation has shaped and been shaped by the world around it. Using time-series plots, geographic maps, and genre-based filters, we aim to create an engaging and insightful experience that brings the history of animation to life.

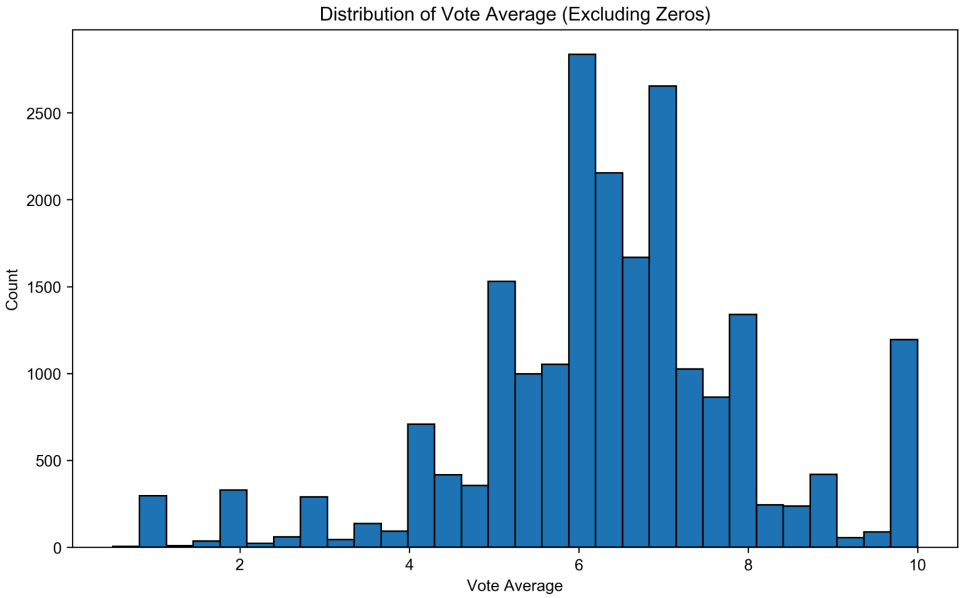
Exploratory Data Analysis

52,000 Animation Movies Dataset

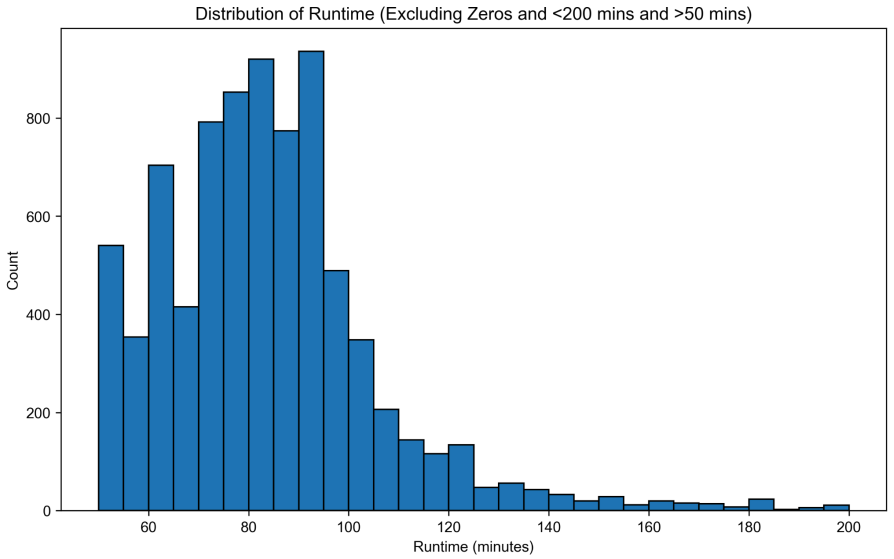
We found the numerical columns to be especially interesting for visualization and therefore analyzed them for both NaN and zero values. Unfortunately, several of these columns, particularly revenue, budget, vote count, and vote average, contained a **significant number of zeros**:

Number of zeros in: vote_average: 30789, vote_count: 30779, revenue: 50845, budget: 50342

For the distribution of vote_average excluding the zeros, we find a **range of 1-10, an average of 6.38 and a median of 6.45.** Furthermore we find that most movies are rated between 4-8 as well as a peak at the far right of the distribution of movies rated with 10.0.

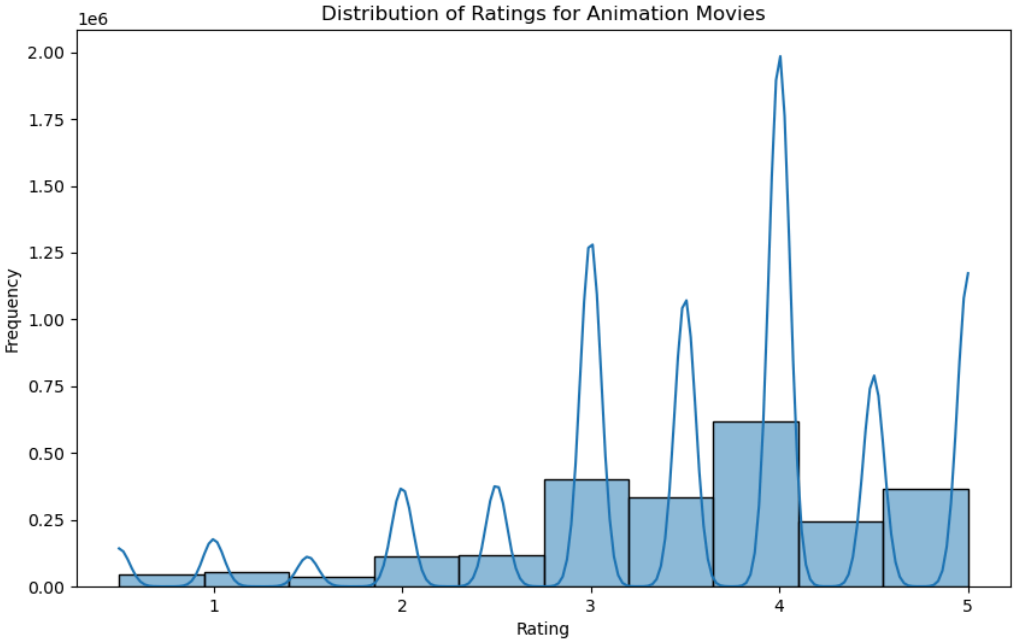


After analyzing movie **runtimes**, we discovered that many films fall into the short movie category, with thousands running for less than 10 minutes.To focus on movies for our visualizations, we restricted our dataset to films with runtimes between 50 and 200 minutes, which left us with **8,062 movies** having an **average runtime of 82.9 minutes**.

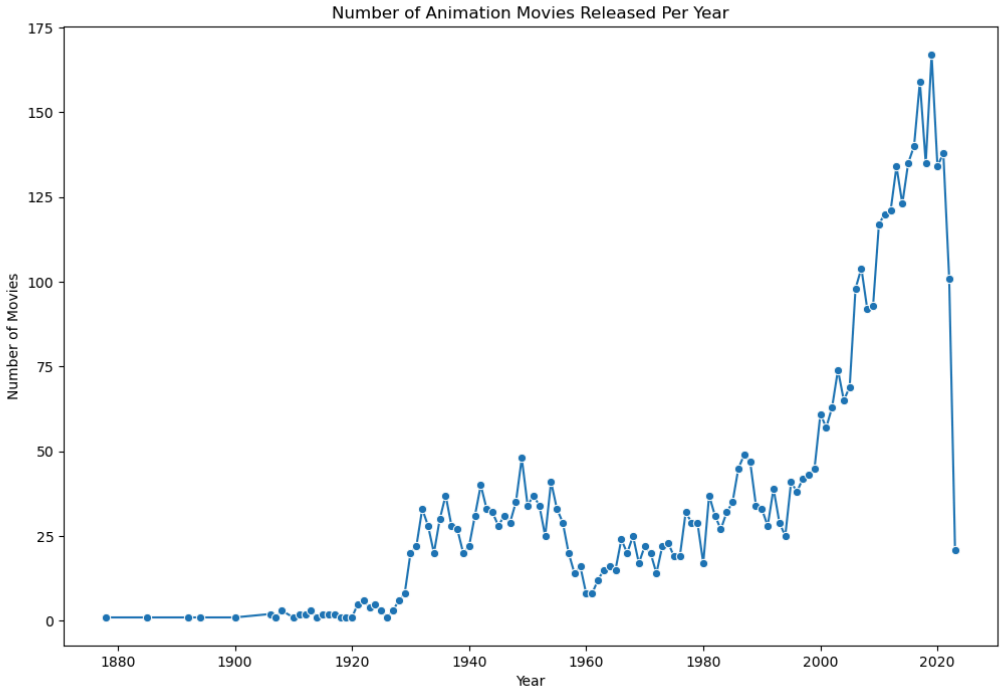


MovieLens Dataset:

Preprocessing steps included filtering the `movies.csv` file by the "Animation" genre, extracting release years using regex, and merging with the `ratings` and `tags` files. We dropped missing values from `tags.csv` and `links.csv` to ensure consistency. The dataset is clean and well-structured. Basic statistics show an average rating of **3.54** (out of 5), with most ratings clustered around **3-4.5**.



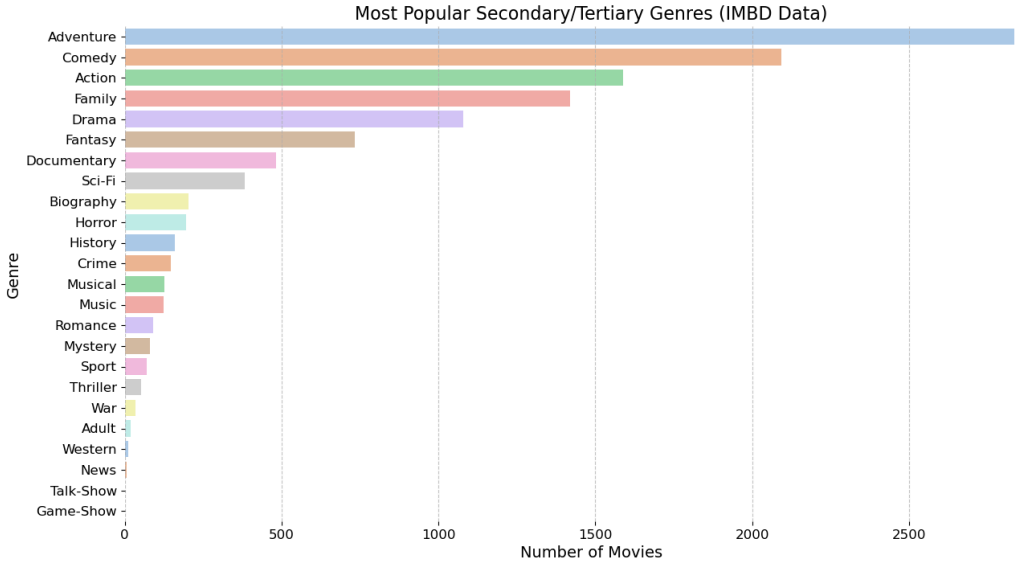
We also visualized the yearly number of animation movie releases, revealing steady growth and notable peaks. These insights help frame how audience interest and production have evolved over time.

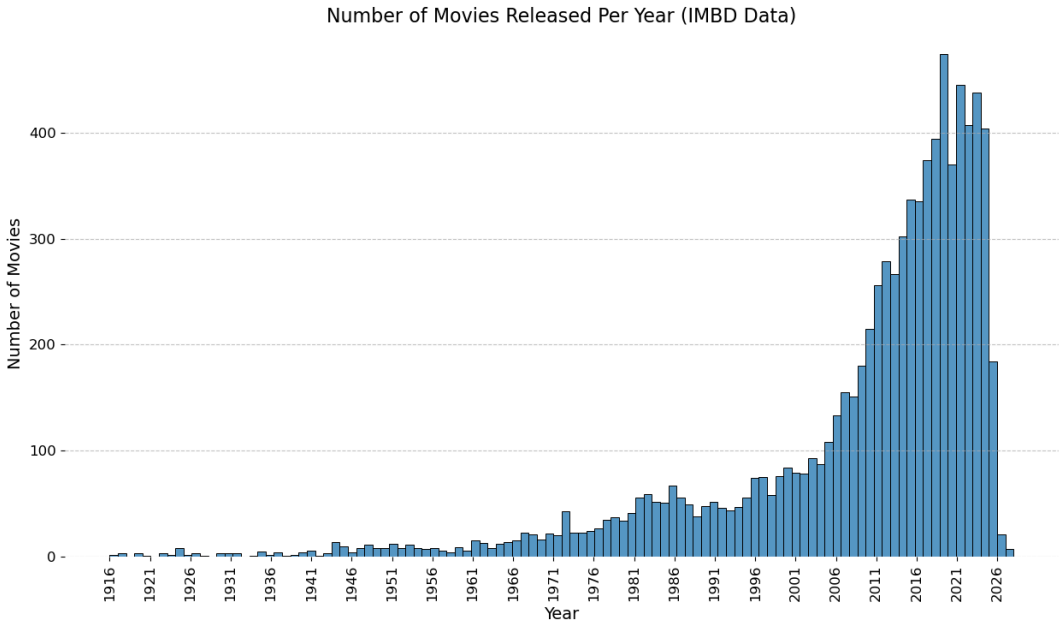


IMBD Dataset:

After filtering by animations and movies, we obtain a total of 10,000 animated movies to analyze in the IMBD dataset. Of these 10,000, approximately **20%** of the dataset has either **missing year of release or movie length**. Furthermore, only **~55%** of them have **published ratings**. Below we see some basic plots describing simple, yet relevant characteristics of the datasets.

The first plot below shows the most popular genres typically paired with animation movies. We find that adventure, comedy, and action are amongst the top three. Based on the second plot, we observe that the quantity of movies released per year increases. We notice this peaks at around **2020** and then begins to decline. An initial explanation for this could be the coinciding **COVID pandemic**.

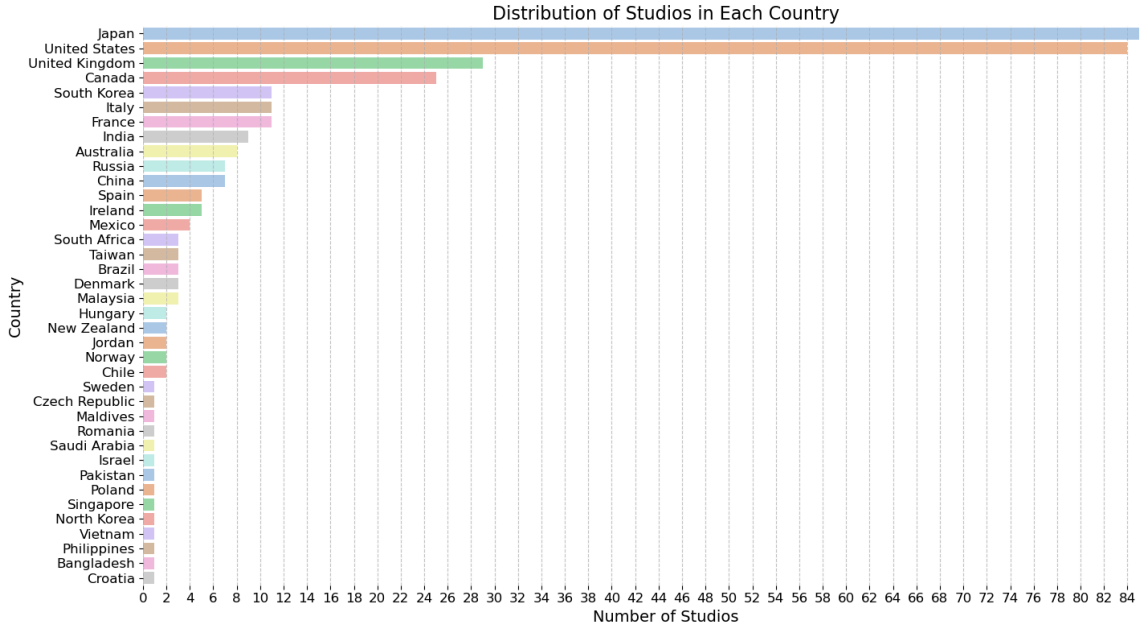


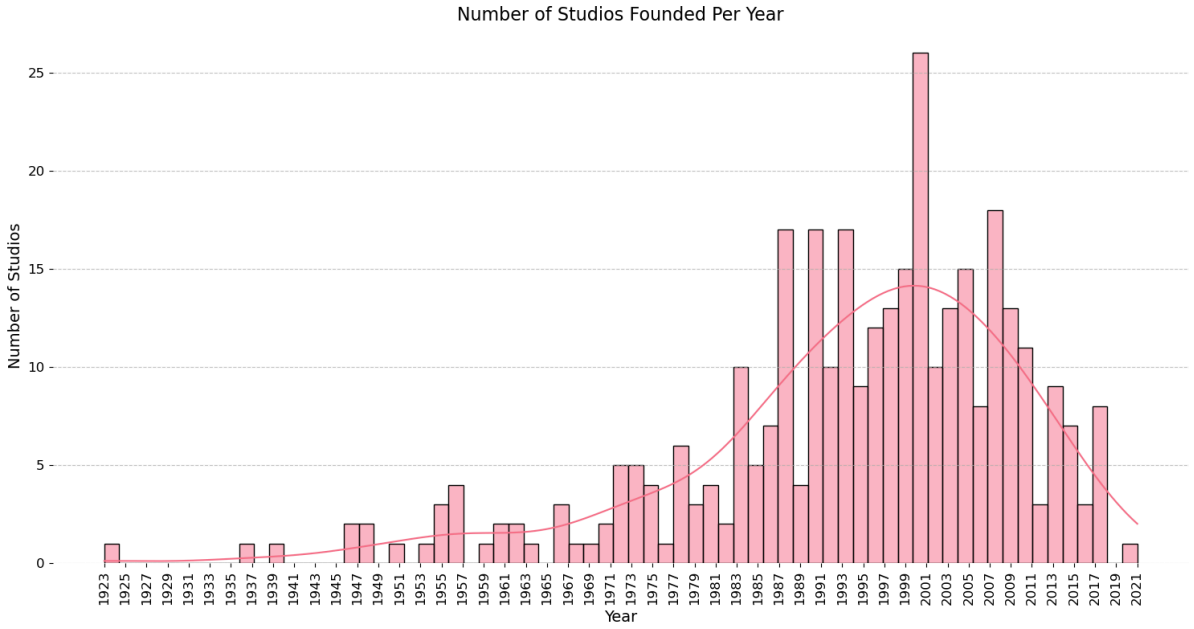


Animation Studio Dataset:

We observe that **Japan and the United States** are top contenders for having the most animation studios. This is expected since both countries are known for their high quality animation production.

Furthermore, we see the distribution formed by the number of studios founded per year is approximately gaussian. The peak is around the beginning of the **21st century.**



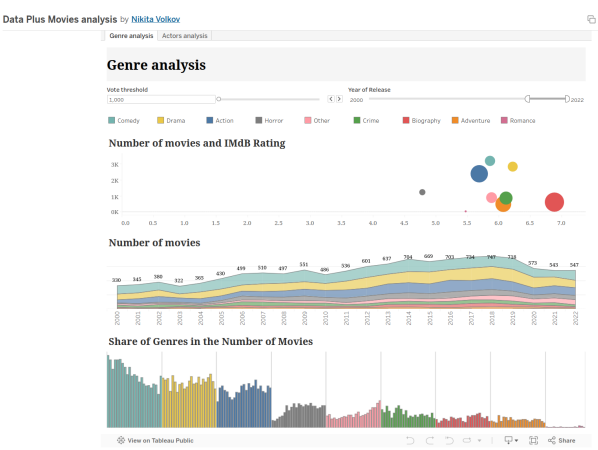


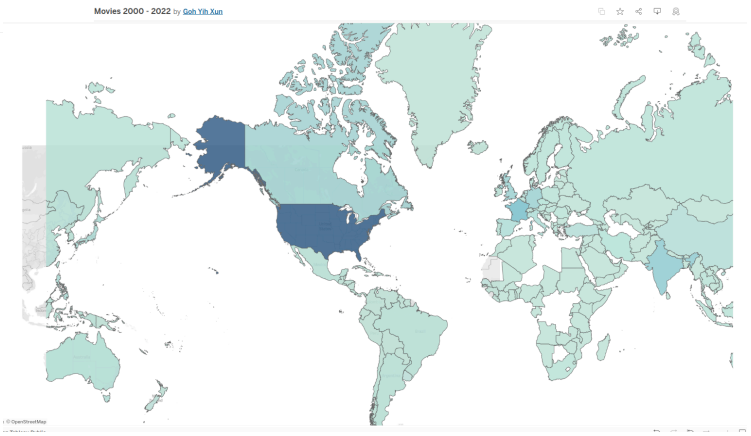
Related Work

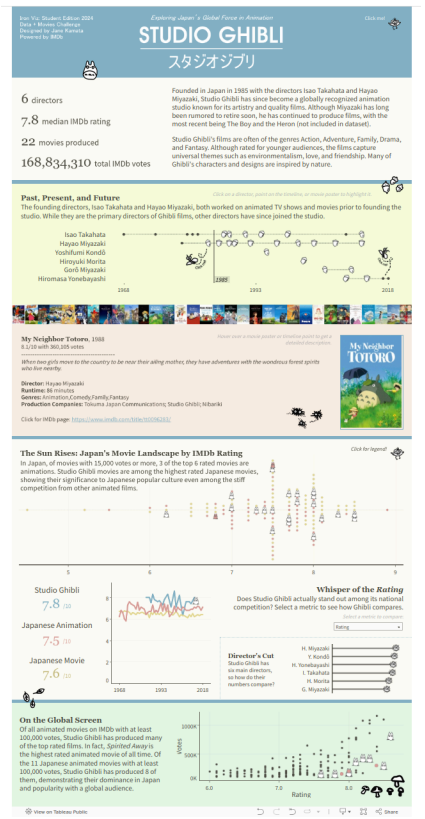
Previous works that utilize IMDb's movie dataset can be found through the Tableau and IMDb collaboration, which presents a wide range of visualizations on topics such as genre popularity, actor networks, or box office performance. While visually engaging, these visualizations **typically focus on a single dimension, like trends over time or comparisons across genres, often lacking interactivity or a broader contextual narrative.**

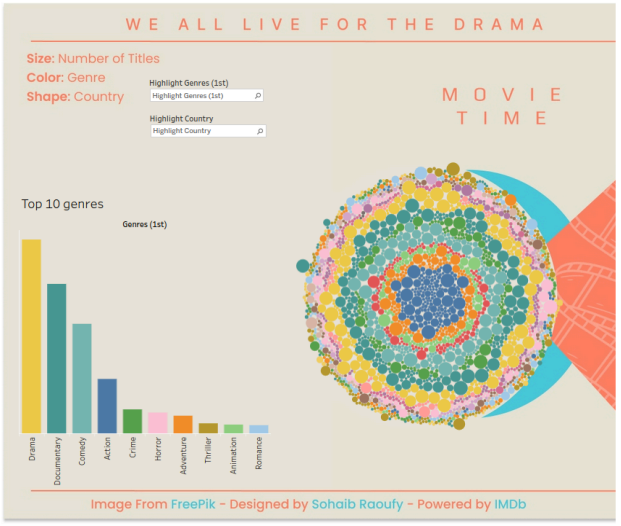
In contrast to these approaches, our project takes a **data-driven and exploratory angle.** Rather than analyzing a single film or focusing narrowly on one aspect, we aim to build an interactive and cohesive visualization that uncovers macro-level trends in animation movies spanning time, geography, and viewer sentiment. Using the MovieLens dataset, which is typically used for recommender systems or general rating analysis, we shift the focus toward storytelling through data: revealing how animated films have evolved globally, how audience ratings reflect changing tastes, and how external events or technological advancements may have influenced production. Our visualization empowers users to filter by genre, rating, or language allowing for a personalized exploration that bridges quantitative insight with cultural context.

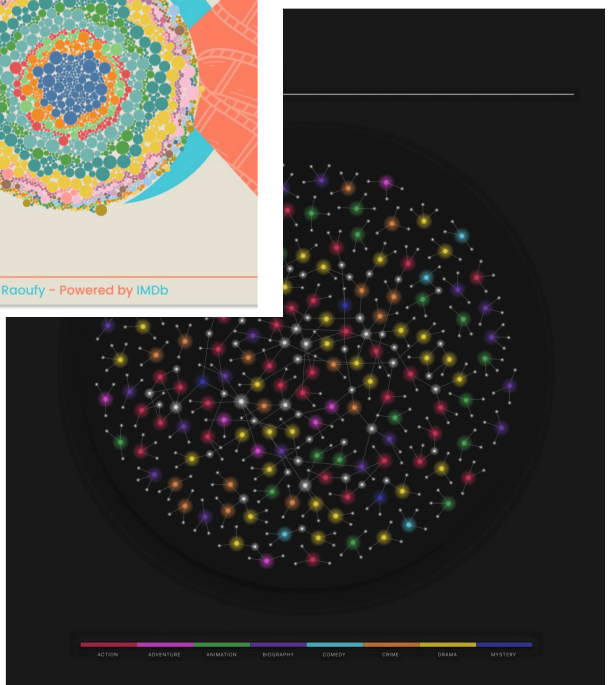
Inspirations











Links for inspirations above:

- https://public.tableau.com/app/profile/nikita.volkov/viz/DataPlusMoviesanalysis/Genreanalysis?_gl=1*1u4pgk6*_ga*MTgwMDcxMzEwNi4xNzQyNDY0MTIy*_ga_8YLN0SNXVS*MTc0MjQ2NDEyMS4xLjEuMTc0MjQ2NjA3OC4wLjAuMA..
- https://public.tableau.com/app/profile/gohyihxun/viz/Movies2000-2022/Map?_gl=1*60lelu*_ga*MTgwMDcxMzEwNi4xNzQyNDY0MTIy*_ga_8YLN0SNXVS*MTc0MjQ2NDEyMS4xLjEuMTc0MjQ2NjA1NC4wLjAuMA..
- https://public.tableau.com/app/profile/jane.kamata/viz/StudioGhibliGlobalForceinAnimationIronVizStudentEdition2024/StudioGhibli?_gl=1*lowgz4*_ga*MTgwMDcxMzEwNi4xNzQyNDY0MTIy*_ga_8YLN0SNXVS*MTc0MjQ2NDEyMS4xLjEuMTc0MjQ2NjQ0Mi4wLjAuMA..
- https://public.tableau.com/app/profile/p.padham/viz/TheMovieNetwork/TheMovieNetwork?_gl=1*139jjqv*_ga*MTgwMDcxMzEwNi4xNzQyNDY0MTIy*_ga_8YLN0SNXVS*MTc0MjQ2NDEyMS4xLjEuMTc0MjQ2NjM3OS4wLjAuMA..
- https://public.tableau.com/app/profile/sahar.sadat/viz/GenreDistributionPerCountry/Dashboard3?_gl=1*10i5w9t*_ga*MTgwMDcxMzEwNi4xNzQyNDY0MTIy*_ga_8YLN0SNXVS*MTc0MjQ2NDEyMS4xLjEuMTc0MjQ2NjM5OC4wLjAuMA..