BIG DATA VISUAL ANALYTICS (CS661)

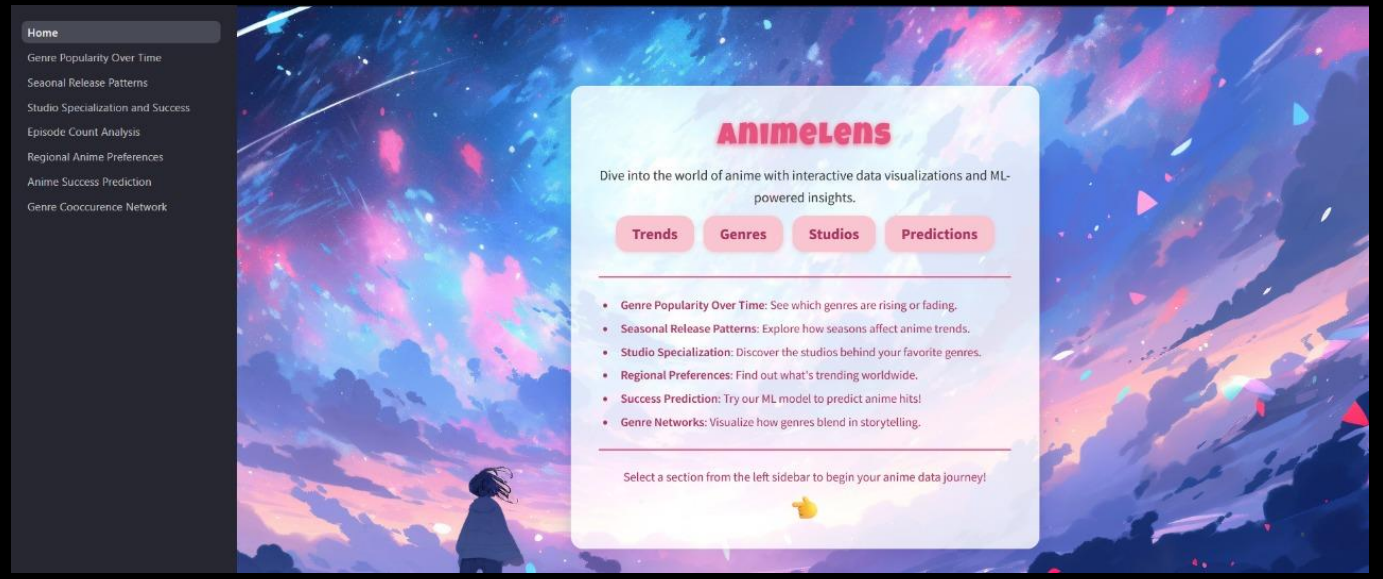
AnimeLens: Visual Analytics System for Exploring Trends, Genres, and Viewer Preferences

INTRODUCTION-:

Our goal is to develop an interactive visual analytics system that uncovers trends and patterns in the anime landscape.

SPECIFIC TASKS-:

1. Genre Popularity Over Time
2. Seasonal Release Patterns
3. Studio Specialization and Success
4. Episode Count vs. Popularity
5. Regional Anime Preferences
6. Success Prediction Model
7. Genre Co-occurrence Network



DATASET DESCRIPTION

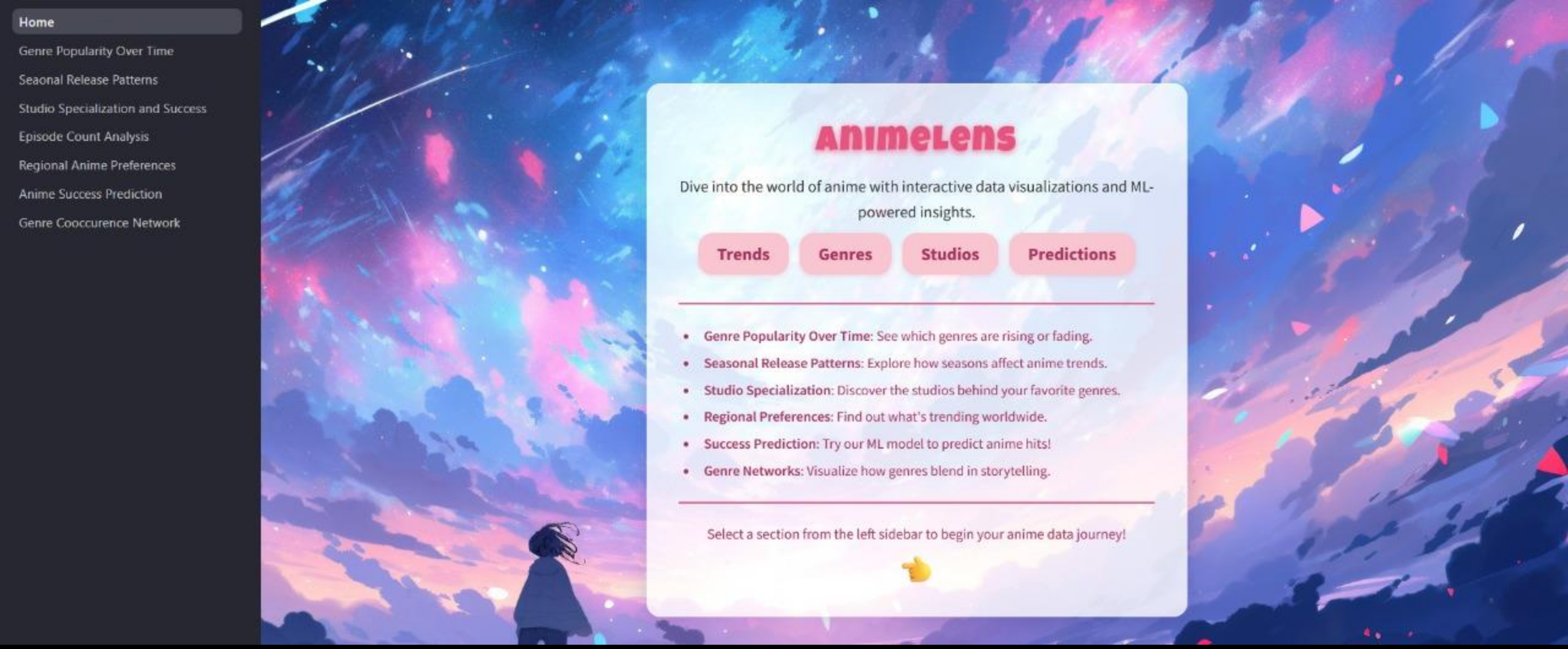
We have used publicly available anime datasets, primarily sourced from Kaggle and the MyAnimeList API. These datasets include information such as-:

- Anime title, type (TV, Movie, OVA, etc.)
- Release year and season
- Number of episodes and average episode duration
- Genre tags
- Average rating (MAL Score)
- Studio name and producer details

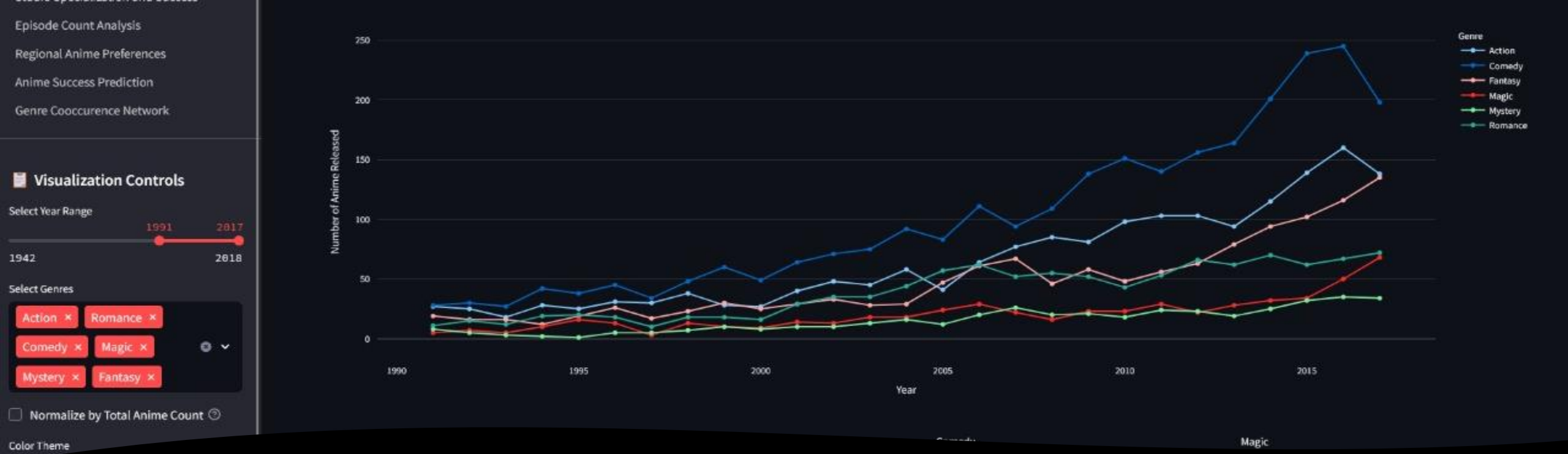
Dataset Link :https://www.kaggle.com/datasets/azathoth42/myanimelist

LIBRARIES USED

- PANDAS - Used for data manipulation, cleaning, and transformation.
- NUMPY-Provided fast numerical operations and array handling, supporting efficient computation for statistics and feature engineering.
- STREAMLIT-Provided the web application framework for the dashboard. Streamlit was selected for its simplicity, rapid development workflow, and seamless integration with Python
- PLOTY EXPRESS-Enabled rapid creation of standard visualizations with minimalcode, making it easy to prototype and iterate on visual analytics modules.
- PLOTY GRAPH OBJECTS-Used for advanced or customized visualizations wherefine-grained control over chart elements was required.
- NETWORKX-Utilized for network analysis and visualization, specifically to constructand display genre co-occurrence graphs
- SCIKIT-LEARN-Used for implementing machine learning algorithms, including modeltraining, validation, and evaluation.



DASHBOARD OVERVIEW



Genre Popularity Over Time-:

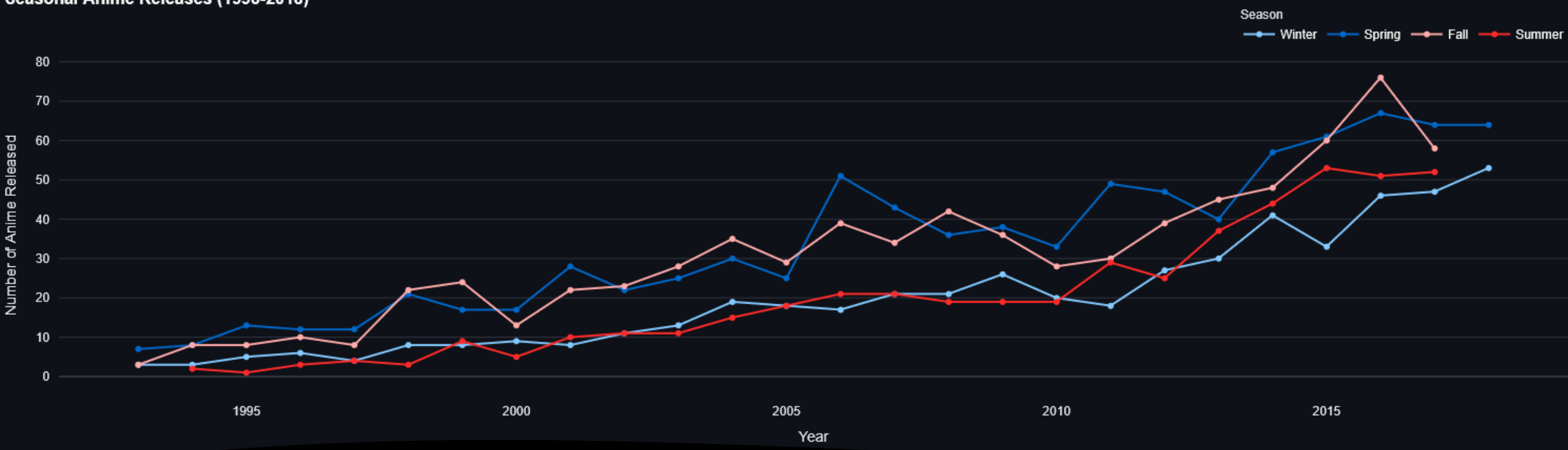
1. Track how different anime genres (Action, Romance, Slice of Life, etc.) evolved in popularity over the years.
2. We have analysed trends, yearly comparison, dominant generes and genre growth.
3. We have also added normalization option

Genre Popularity Over Time:-

1. Tab 1: Genre popularity Trends Over time
    - A line chart is created using **Plotly** to visualize the popularity trends of selected anime genres over time
    - The chart layout is further customized using hovermode . It ensures that hover tooltips display data for all genres at a given year
    - Below the chart, shows statistics like count and growth w.r.t to previous year for a particular anime.

2. Tab 2: Genre comparison over year
    - provides an interactive way to compare anime genre distributions across years, combining visual insights (bar chart) with numerical analysis
    - A grouped bar chart is created where: x-axis = selected years, y-axis = the metric (e.g., percentage/count of anime). Bars are grouped by genre, with colors for genre.
    - Genre change analysis table : percentage changes in the metric between the min and max years are calculated. The results are sorted by percentage change in descending order
3. Tab 3: Popularity Heatmap
    - This tab provides two key visualizations
        1. A heatmap that shows the popularity of anime genres over time, allowing users to identify trends and patterns
        2. A horizontal bar chart that highlights the genres with the most years as dominant, offering insights into long-term popularity.

4. Tab 4: Growth Analysis
    - calculates growth metrics of anime genres over time, displays them in a table, and visualizes the results using bar and bubble charts
    - The CAGR is calculated using the formula: CAGR = (((last_value / first_value) ** (1 / years_diff)) - 1) * 100
    - A horizontal bar chart is created : The y-axis represents genres, and the x-axis represents their CAGR.
    - A bubble chart is created : The x-axis represents the initial value, and the y-axis represents the final value.
    - The bubble size corresponds to the absolute change in value, and the bubbles are colored by genre. The chart includes hover tooltips and labels for better interactivity.
    - A reference line y=x is created if bubble of a particular genre is above it it means it has growth in the final year w.r.t initial year



Seasonal Release Patterns:

1. Analyze the distribution of anime releases by season (Winter, Spring, Summer, Fall)
2. We have variables like years and seasons, which we can modify to get results according to our requirements.

Seasonal Release Patterns :

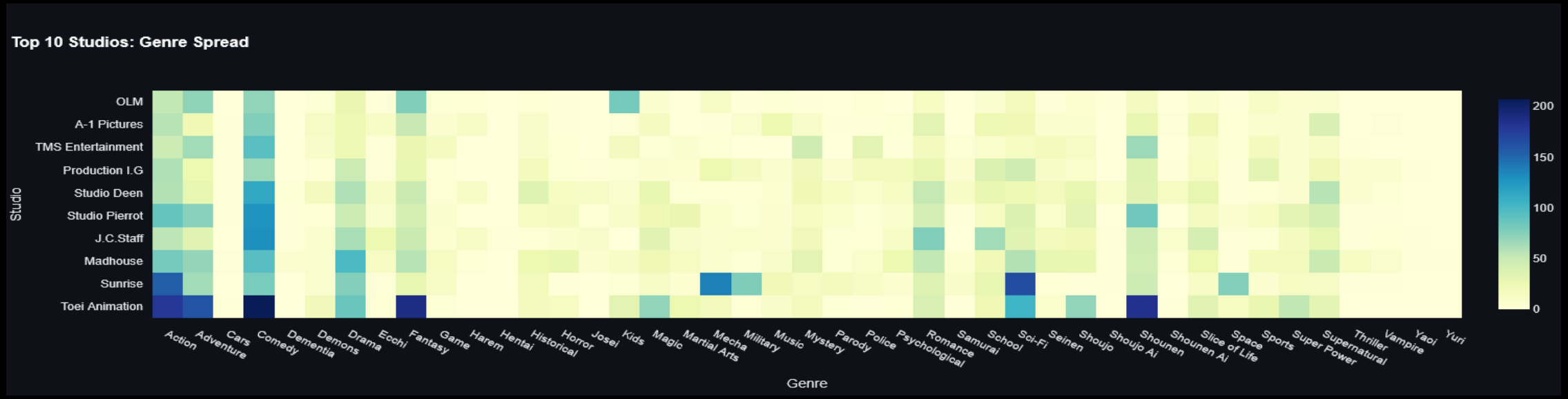
1. Tab 1: Release Trends Over Time
    - A time-series line chart is generated using Plotly Express to visualize the number or percentage of anime releases by season across years.
    - Option to normalize by year shows each season's share as a percentage of total yearly releases
2. Tab 2: Ratings Analysis
    - Bar charts display season wise average anime score and average popularity rank
    - Line Charts show the year-wise average anime score and average popularity rank
3. Tab 3: Seasonal Heatmap

    •Heatmaps are created to visualize the count of anime releases, average scores and average popularity ranking by season and year, with color intensity representing volume.
1. Tab 4:Comparative View
    - Box plots are used to visualize the distribution of anime scores and average popularity raanking for each season

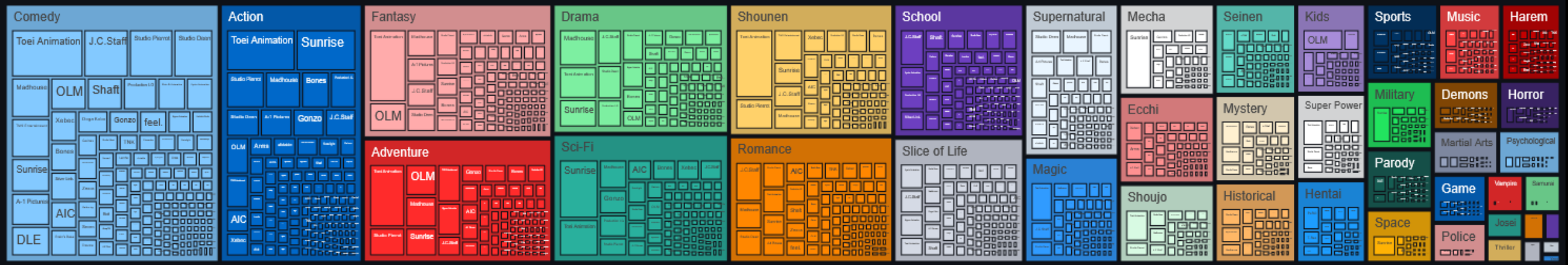
Studio Specialization and Success :

1. Treemap: Studio Dominance by Genre
    - The treemap is implemented using Plotly Express's px.treemap, providing a hierarchical visualization where each block represents a genre and each rectangle within shows a studio's production volume in that genre
    - This chart is ideal for quickly assessing which studios are most influential within specific genres, revealing concentration or diversity at a glance

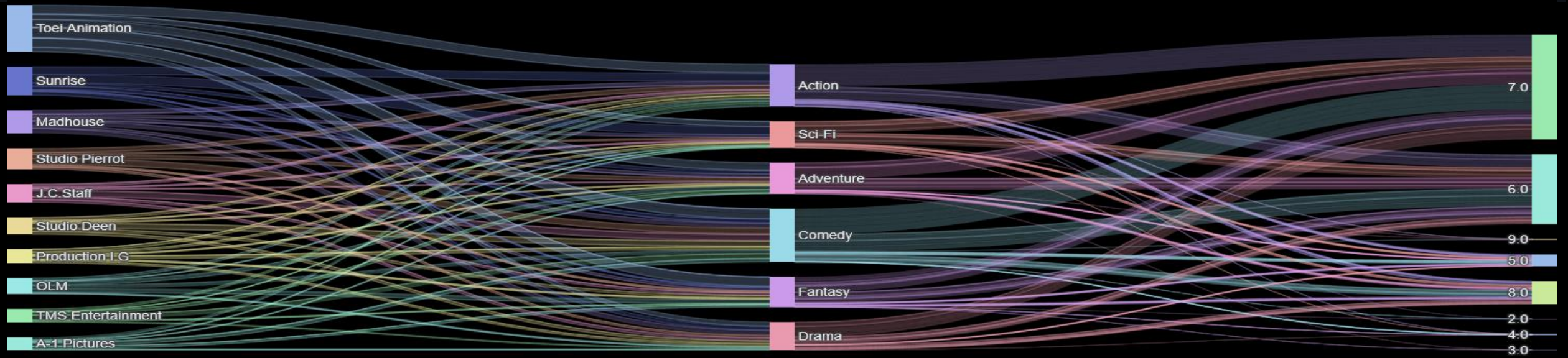
2. Sankey Diagram: Studio → Genre → Rating
    - The Sankey diagram is created using Plotly's go.Sankey, visualizing the flow of anime counts from studios to genres to average ratings, with flow width proportional to volume and color encoding ratings
    - This visualization is valuable for understanding how studio output distributes across genres and translates into quality, highlighting studios that consistently produce highly-rated anime in particular genres.

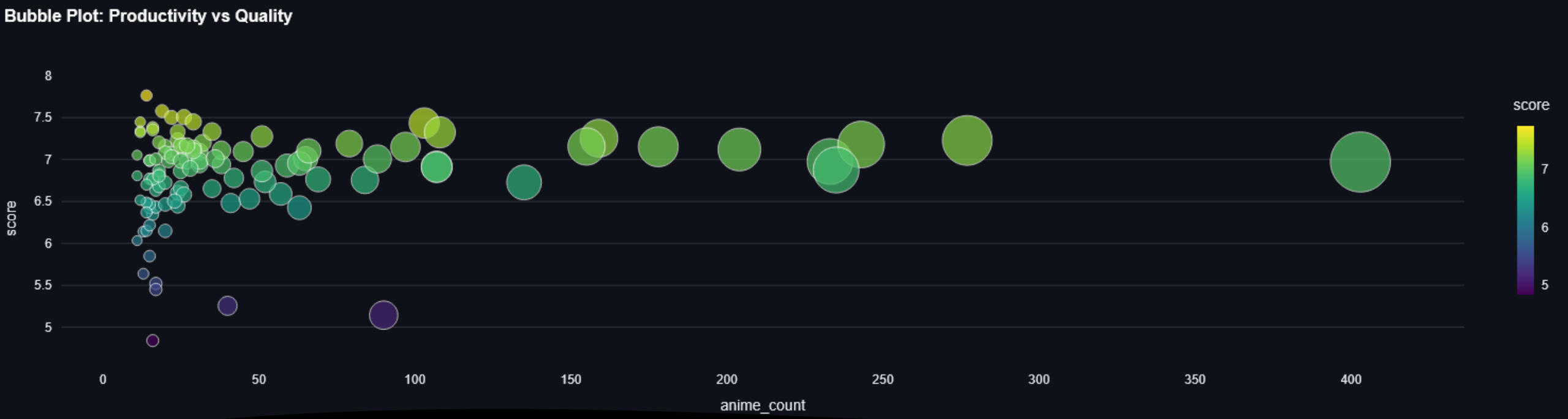


Treemap: Studio Dominance by Genre



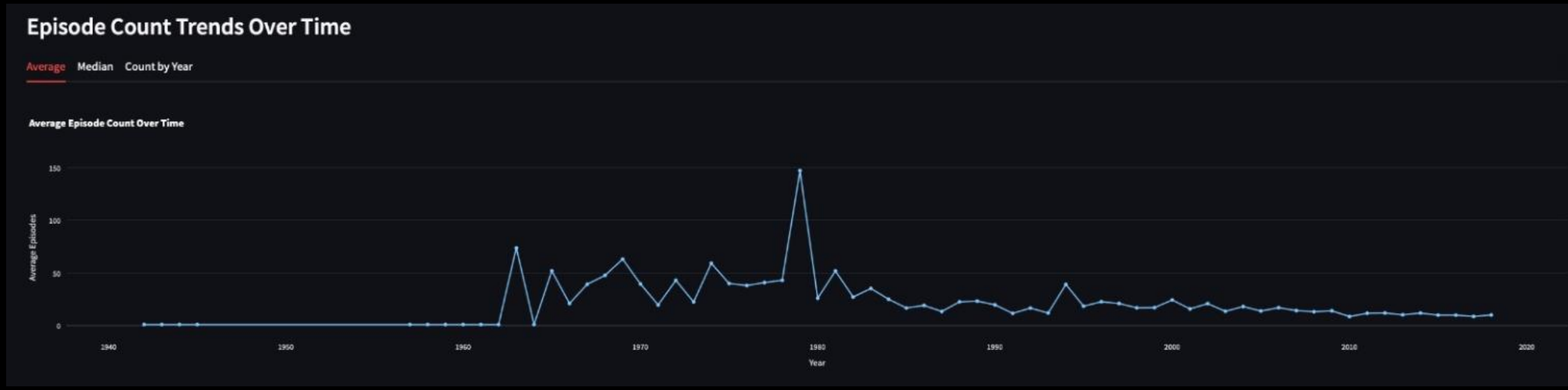
Studio → Genre → Rating





Bubble Plot – Studio Productivity vs Quality

- A. bubble plot is implemented using Plotly Express's pxscatter, where each point represents a studio, the x-axis shows the number of anime produced, the y-axis displays average score, and bubble size encodes productivity.
- This visualization is highly interactive: when you hover over a bubble, detailed information about the studio (name, anime count, score) appears instantly, making it easy to identify top-performing studios at a glance.
- Bubble plots are ideal for comparing multiple metrics simultaneously and spotting outliers or clusters, with Plotly ensuring smooth zoom, pan, and tooltip functionality for deeper data exploration.



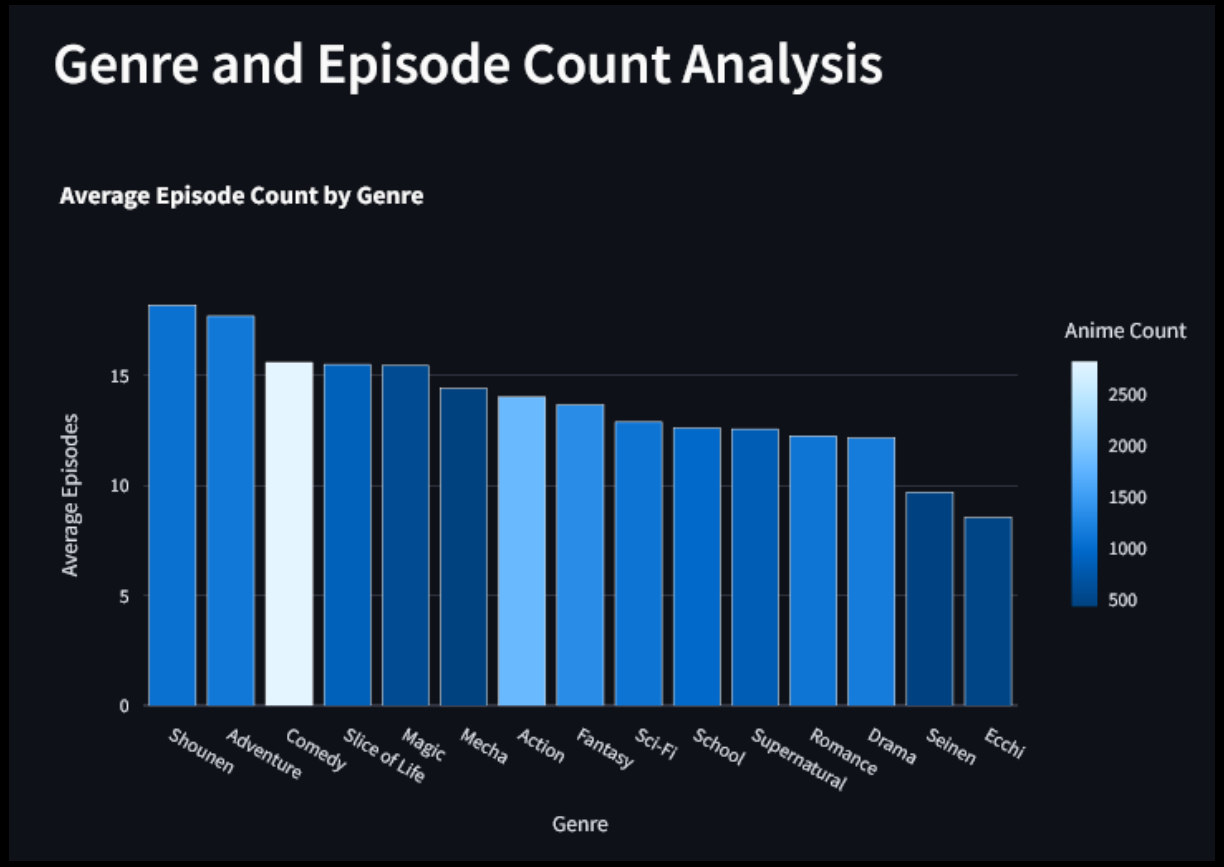
Episode Count vs. Popularity :

1. Investigate whether long-running anime are more popular than shorter seasonal shows
2. Study whether shorter anime tend to receive higher ratings due to tighter storytelling.

Episode Count vs. Popularity -:

1. Tab 1: Average Episode Count Over Time (Line Chart)
    - The line chart tracks the average episode count of anime released each year, revealing trends and fluctuations in episode production from 1940 to 2020 (changeable).
    - Notable spikes and declines are easily visible, helping users identify periods of unusually high or low average episode counts in the industry's history
    - The chart layout is further customized using hovermode . It ensures that hover tooltips display data for all genres at a given year

2. Tab 2: Average Episode Count by Genre
    - A bar chart is created using Plotly Express (px.bar) to visualize the average episode count for each major genre, helping users compare which genres tend to have longer or shorter series.
    - The chart layout is customized using fig.update_layout(hovermode="x unified"), ensuring that when hovering, tooltips display the average episode count for all genres at once
    - Data for the bar chart is prepared by grouping the DataFrame by genre and calculating the mean episode count using groupby('genre').agg(avg_episodes=('episodes', 'mean'))
    - The chart layout is customized for interactivity, so hovering over each bar instantly shows the exact average episode count and genre details.

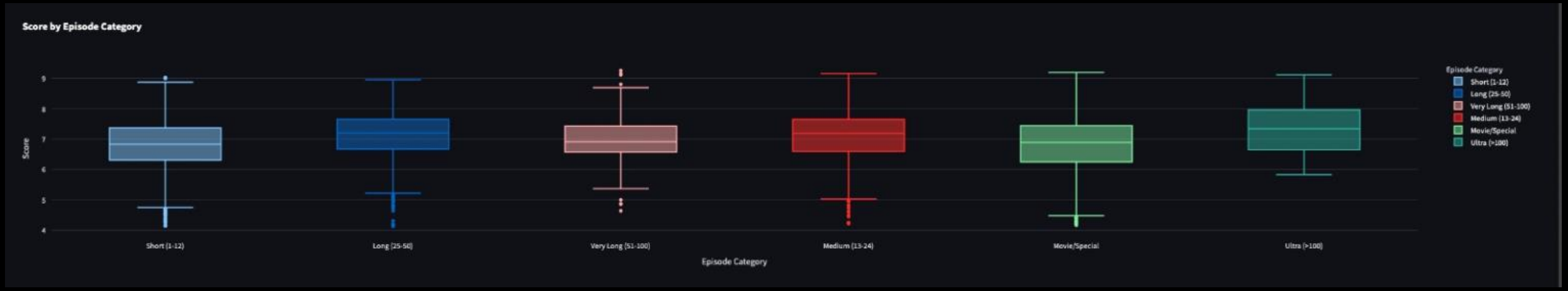


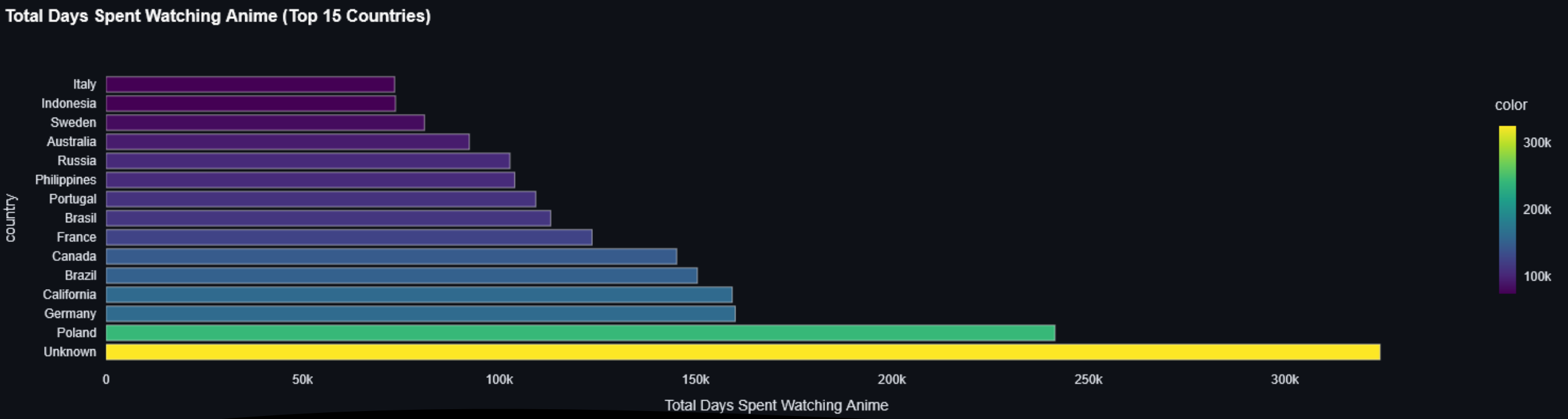
Tab 3: Episode Count Distribution by Genre

1. This box plot(px.box) visualizes the distribution of anime scores across different episode categories (Short, Long, Very Long, Medium, Movie/Special, Ultra), making it easy to compare median scores and variability for each group

2. The plot uses color-coding for each episode category and displays statistical features such as median, quartiles, and outliers, helping users identify which episode lengths tend to receive higher or lower ratings

3. Interactive features (typical in Plotly) allow users to hover over each box to see detailed statistics, supporting deeper exploration of how episode count relates to anime quality

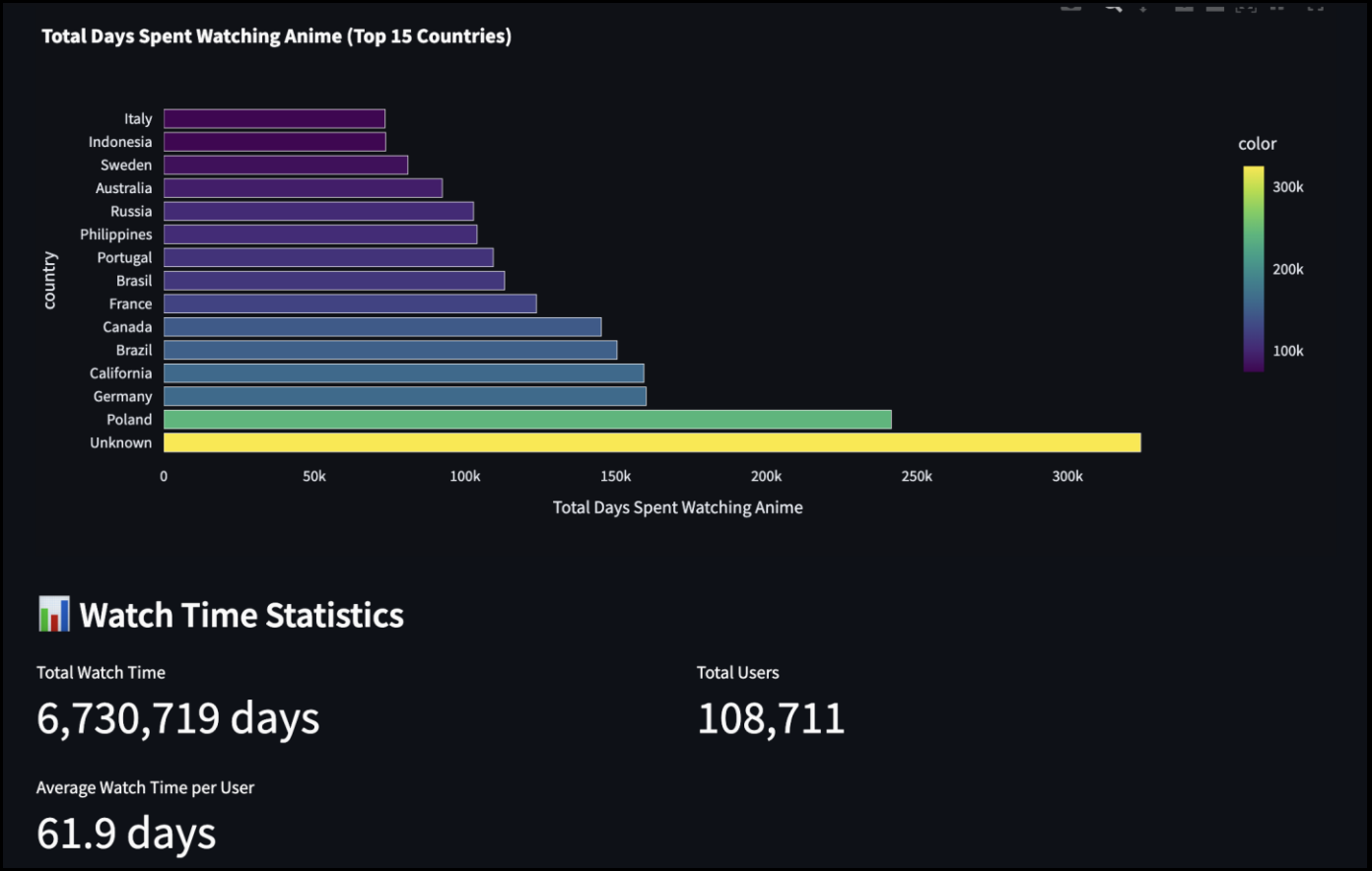




Regional Anime Preferences :

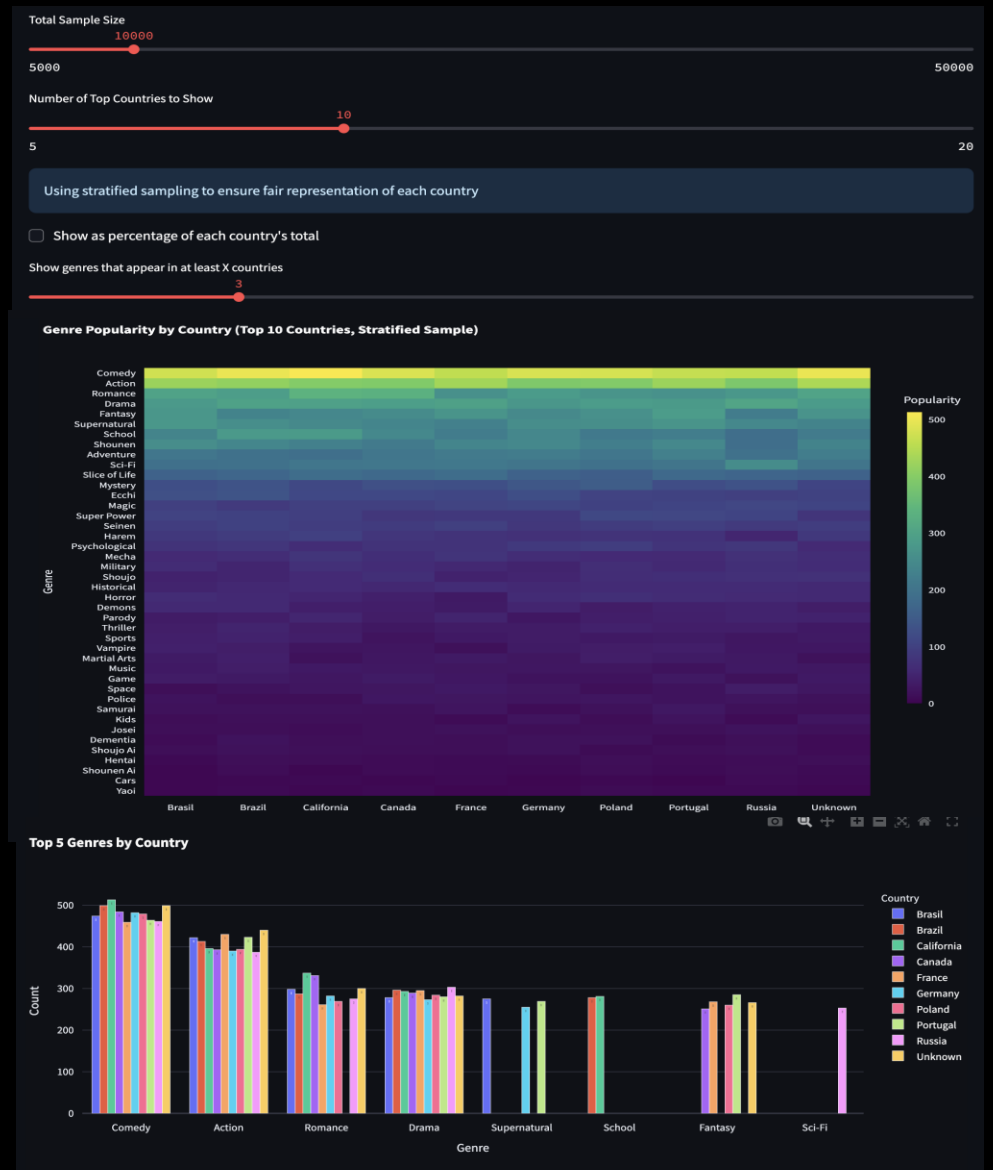
1. Determine region-wise preferences in anime consumption.
2. Analyze watch-time patterns across different countries and genre popularity in various regions.

Tab 1: Watch time analysis



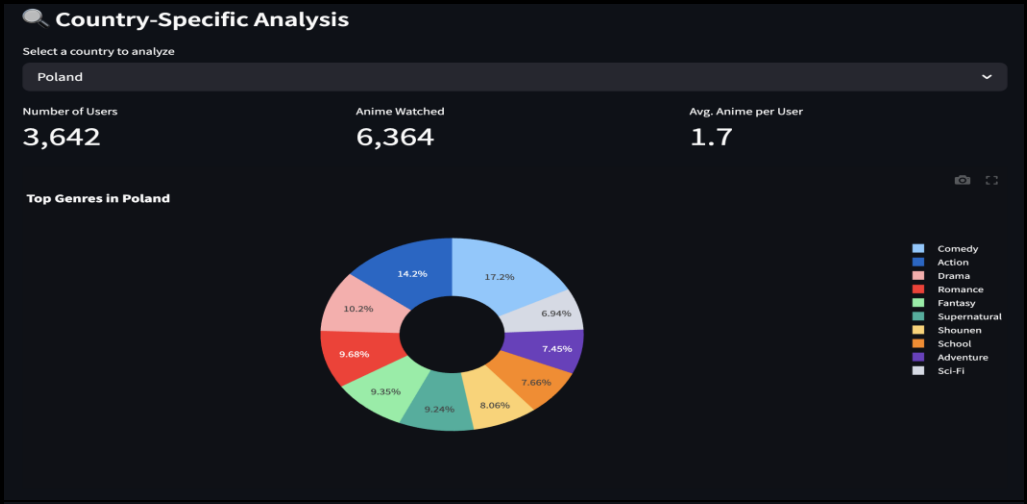
- **Bar Chart:** Visualizes the top 15 countries by total watch time, highlighting regional engagement.
- A horizontal bar chart is created using **Plotly Express** (px.bar()):
  The x-axis represents the total days spent watching anime (watchtime_region.values). The y-axis lists the top 15 countries
- **Watch Time Statistics**: Displays total and average watch time, as well as the number of users represented in the dataset.

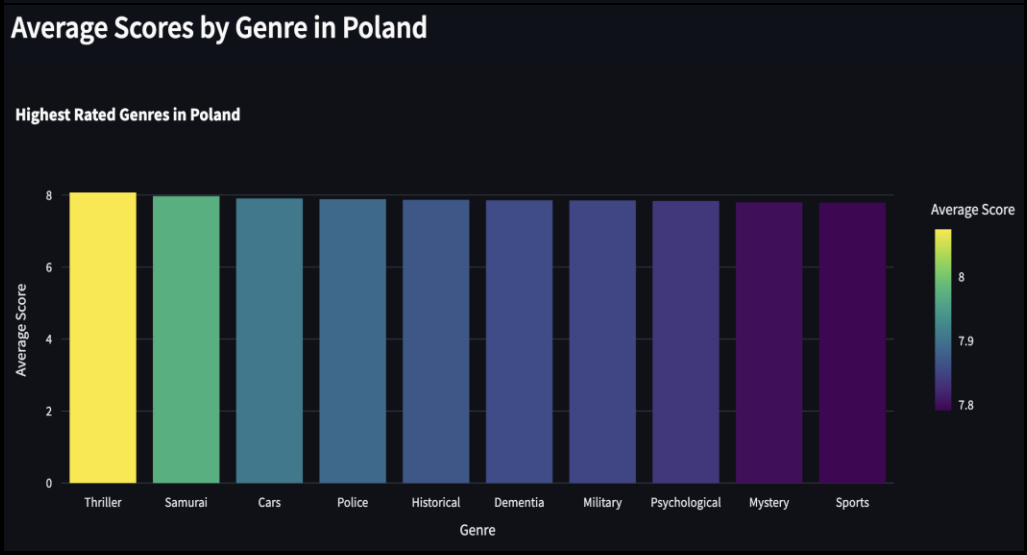
Tab 2: Genre Preferences



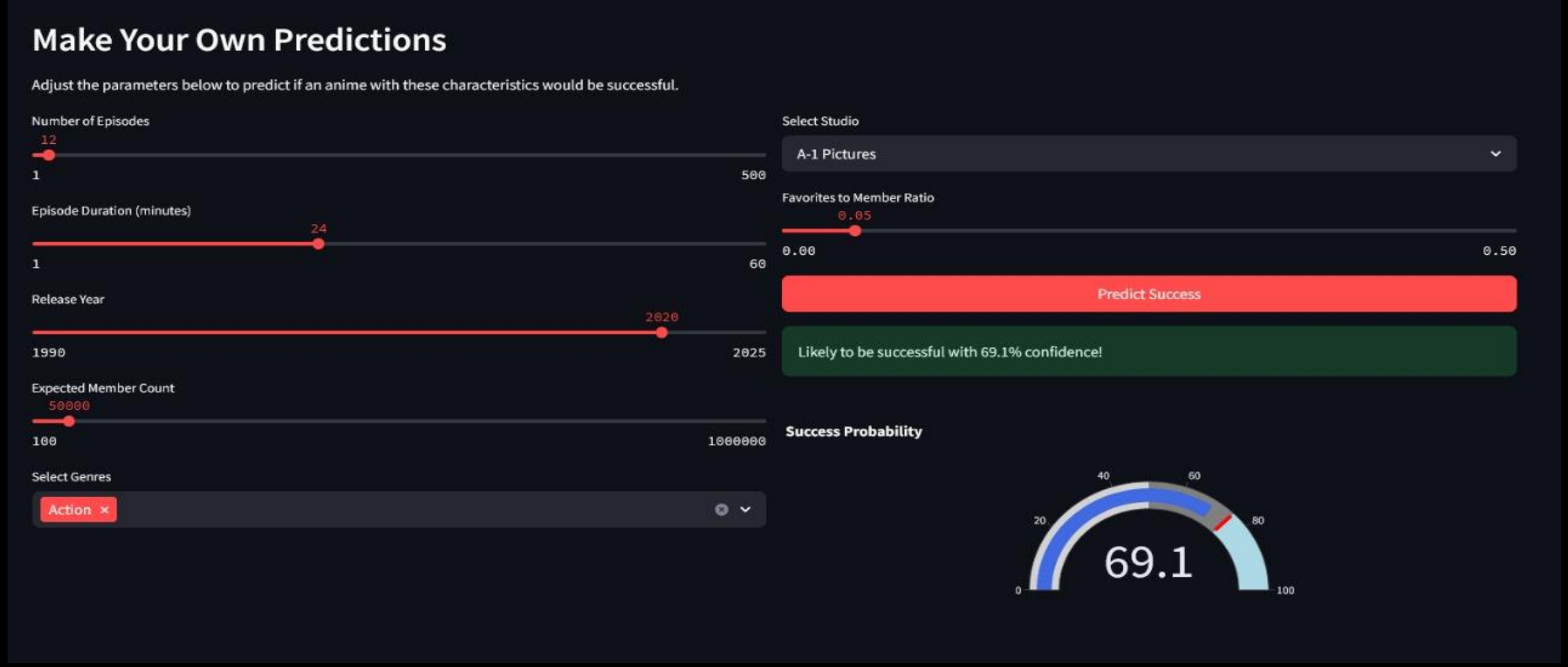
- Using stratified sampling to ensure fair representation of each country
- Slider 1: to set the Sampling Data size ranging from 5,000 to 50,000, with a default value of 10,000.
- Slider 2: lets users select the number of top countries to display, ranging from 5 to 20, with a default value of 10.
- The stratified_sample function : For each country, if the number of records exceeds the maximum allowed , a random sample is taken. Otherwise, all records are included.
- The maximum number of records per country is calculated as sample_size // len(top_regions), ensuring the total sample size matches the user-defined value.
- **Heatmap**: Visualizes genre popularity across countries, with options for normalization and filtering.
- **Top Genres by Country**: Highlights the most popular genres for each country using a grouped bar chart.

Tab 3 : Detailed country specific analysis





- **Country-Level Metrics**: Displays the number of users, anime watched, and average anime per user.
- **Genre Preferences**: Visualizes the top 10 genres in the selected country using a pie chart.
- **Average Scores by Genre**: Highlights the highest-rated genres in the selected country using a bar chart.
- The data is grouped by `genre`, and the average score (`mean`) and count of anime are calculated. Genres with fewer than 5 anime are filtered out to ensure meaningful results.



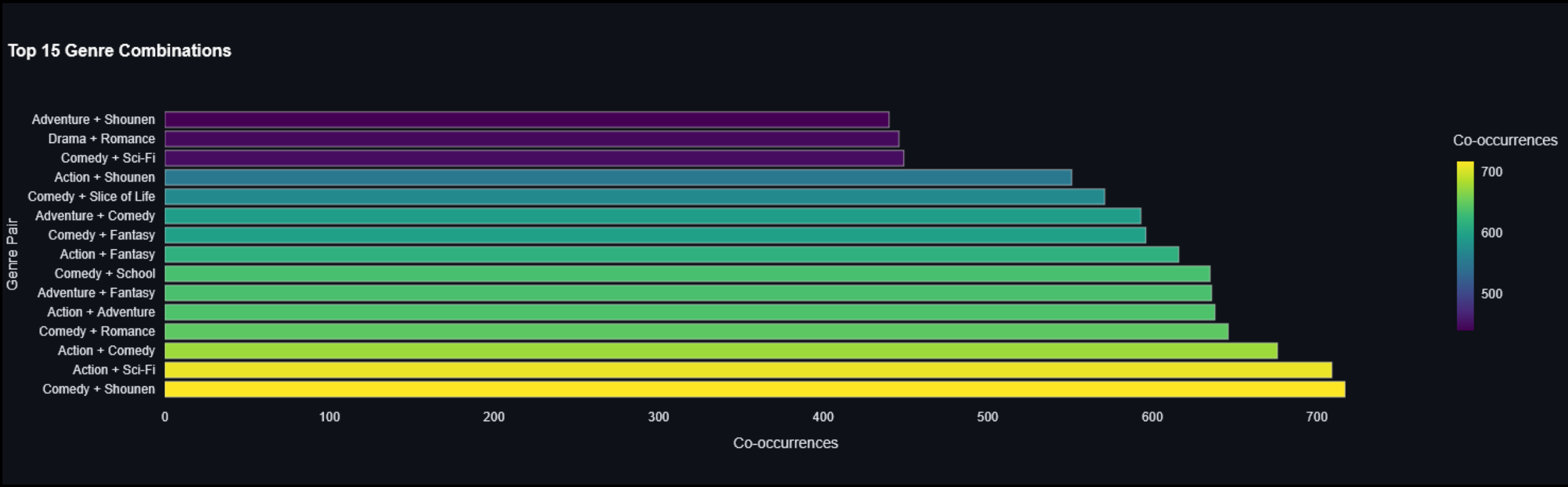
Success Prediction Model :

1. Develop a machine learning model to predict an anime's success based on factors such as genre, episode count, animation studio, and user reviews.

Success Prediction Model :-

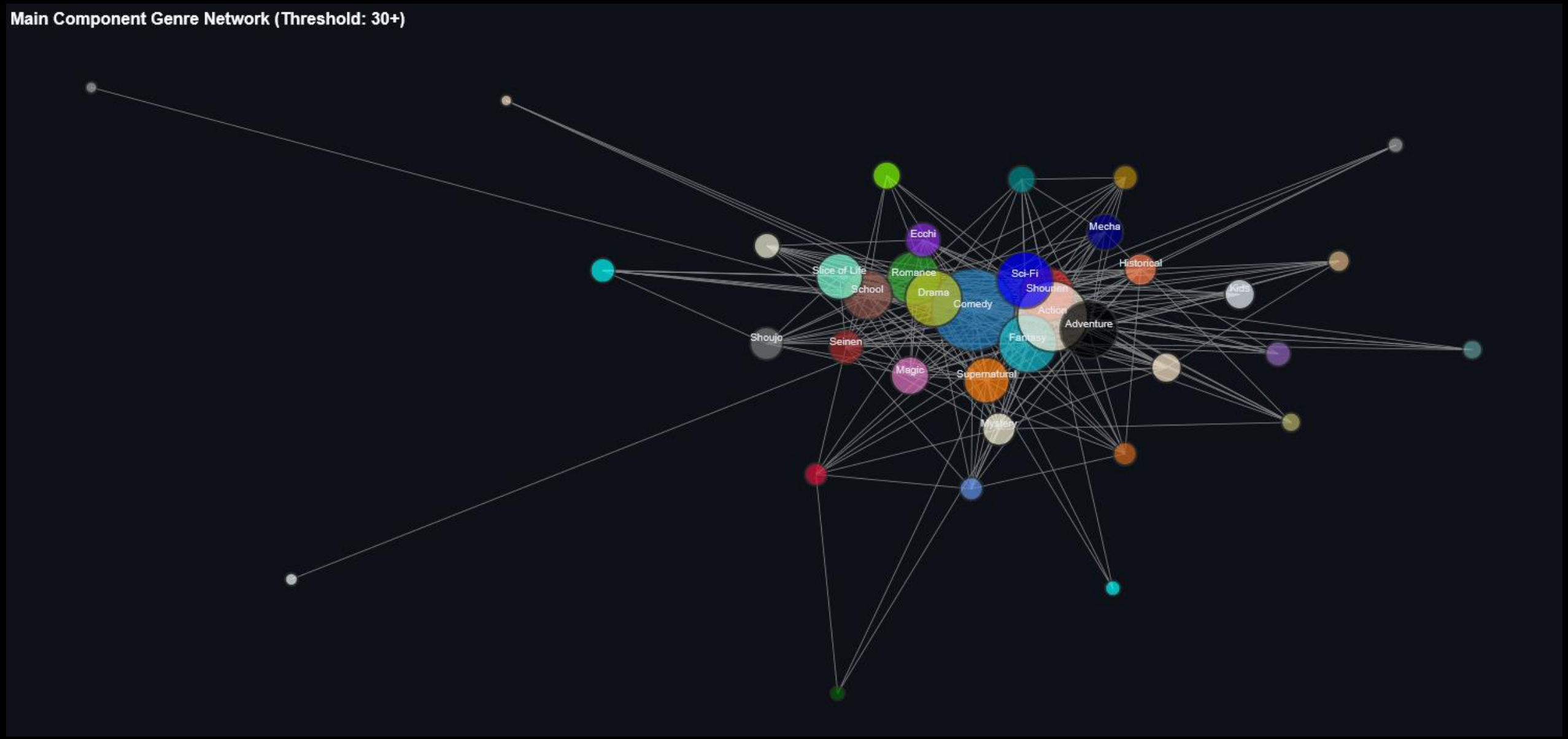
- Tab 1: Model Performance
  - The classification report is displayed to show precision, recall, accuracy and F1-scores.
  - **A confusion matrix is visualized as a heatmap using Plotly to display actual prediction counts.**
- Tab 2: Make Predictions
  - Users can adjust anime features through interactive widgets to input episode count, duration, release year, expected memberes and genres.
  - We have used Random Forest Classifier to make predictions.
  - We have defined a score of >= 6.5 as success.
  - Prediction results are displayed based on the model's prediction with success probability percentage.
- Tab 3 : Feature Importance

  Top 15 important features are displayed along with correlations of features calculated with respect success.



Genre Co-occurrence Network:

1. Create a co-occurrence graph to reveal which genres tend to appear together (e.g., Romance + Comedy, Fantasy + Adventure).
2. Understand popular formulae in anime storytelling.


Main Component Genre Network (Threshold: 30+)
Ecchi
Mecha
Historical
Slice of Life
Romance
Sci-Fi
School
Drama
Shounen
Kids
Comedy
Action
Adventure
Shoujo
Seinen
Fantasy
Magic
Supernatural
Mystery

Genre Co-occurrence Network:-

Build data: It processes raw data to identify significant co-occurrences, filters out less meaningful connections, and builds a graph representation that can be used for visualization or further analysis. The use of Counter, combinations, and NetworkX ensures efficient computation and flexibility in handling the data.

- Tab 1: Co-occurrence visualization
  - Slider - Fix threshold to determines the minimum number of anime in which two genres must co-occur for their connection to be displayed in the visualization
  - **Network Graph**:
    1. displays genres as nodes and their co-occurrences as edges in a graph.
    2. The nx.spring_layout() function generates positions for nodes in the graph using a force-directed algorithm- that arrange nodes (vertices) and edges in a way that minimizes edge crossings and creates an aesthetically pleasing layout
  - **Heatmap Matrix**:
    1. An adjacency matrix is created where rows and columns represent genres, and cell values represent co-occurrence counts. The matrix is filled by iterating over all pairs of genres in the graph
    2. A heatmap is generated using Plotly Express's px.imshow():The x and y axes represent genres. The color intensity represents co-occurrence count
- Tab 2: Genre analytics
  - First chart: A horizontal bar chart is created using Plotly Express (px.bar()) to visualize the top 15 genres by the frequency of their occurrence.
  - Second chart: A horizontal bar chart is created to visualize the top 15 genres by by the number of connections (degree)
  - Analysis of selected genre: If a genre is selected and exists in the graph:
  - The G.neighbors() method retrieves all genres connected to the selected genre.
  - The co-occurrence count for each connection is extracted from the graph's edge attributes (G[selected_genre][n]['weight']).
  - The connections are sorted in descending order of co-occurrence count and Visualised using bar graph
- Tab 3 : Top Combinations
  - This tab provides a comprehensive analysis of genre combinations
    1. **Bar Chart**: Highlights the top 15 genre pairs by co-occurrence count.
    2. **Data Table**: Displays all genre combinations for detailed exploration.
    3. **Download Option**: Enables users to export the data for further analysis.

THANK YOU