ANIME SCORE ANALYSIS

THE POWER OF DATA PREPARATION
THROUGH DATA STORYTELLING

Presented by

Group 1

PART 1

DATA STORYTELLING

ACT 1

DATA SITUATION

1. BUSINESS CONTEXT

BUSINESS CONTEXT

Role

Anime Producer deciding which new project to greenlight.

Goal

Identify factors that strongly influence anime Score.

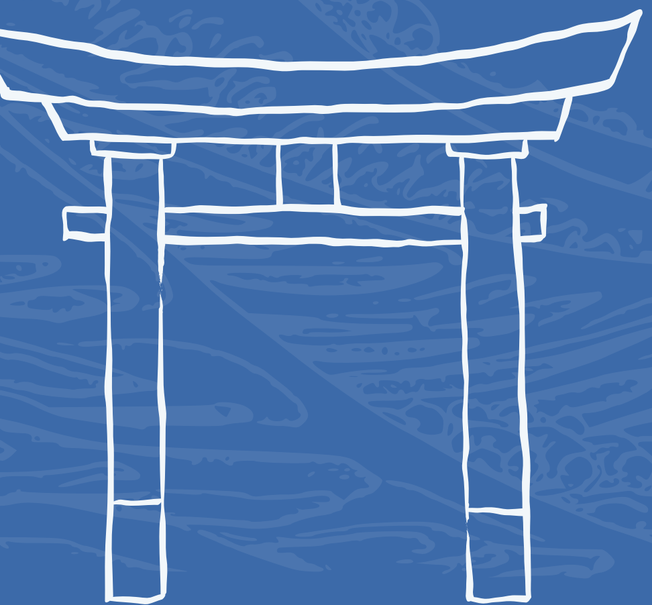
Problem

Decisions depend on reliable insights—but the dataset is messy and misleading.

2. THE CONFLICT

THE CONFLICT

Raw Dataset Issues



- **Hidden Missing, and Contradictory Values**
- **Inconsistent, Fragmented, and Misclassified Metadata**
- **Unreliable Distributions and Identity Ambiguities**

THE CONFLICT

Hidden Missing, and Contradictory Values

Large portions of the dataset pretends to exist

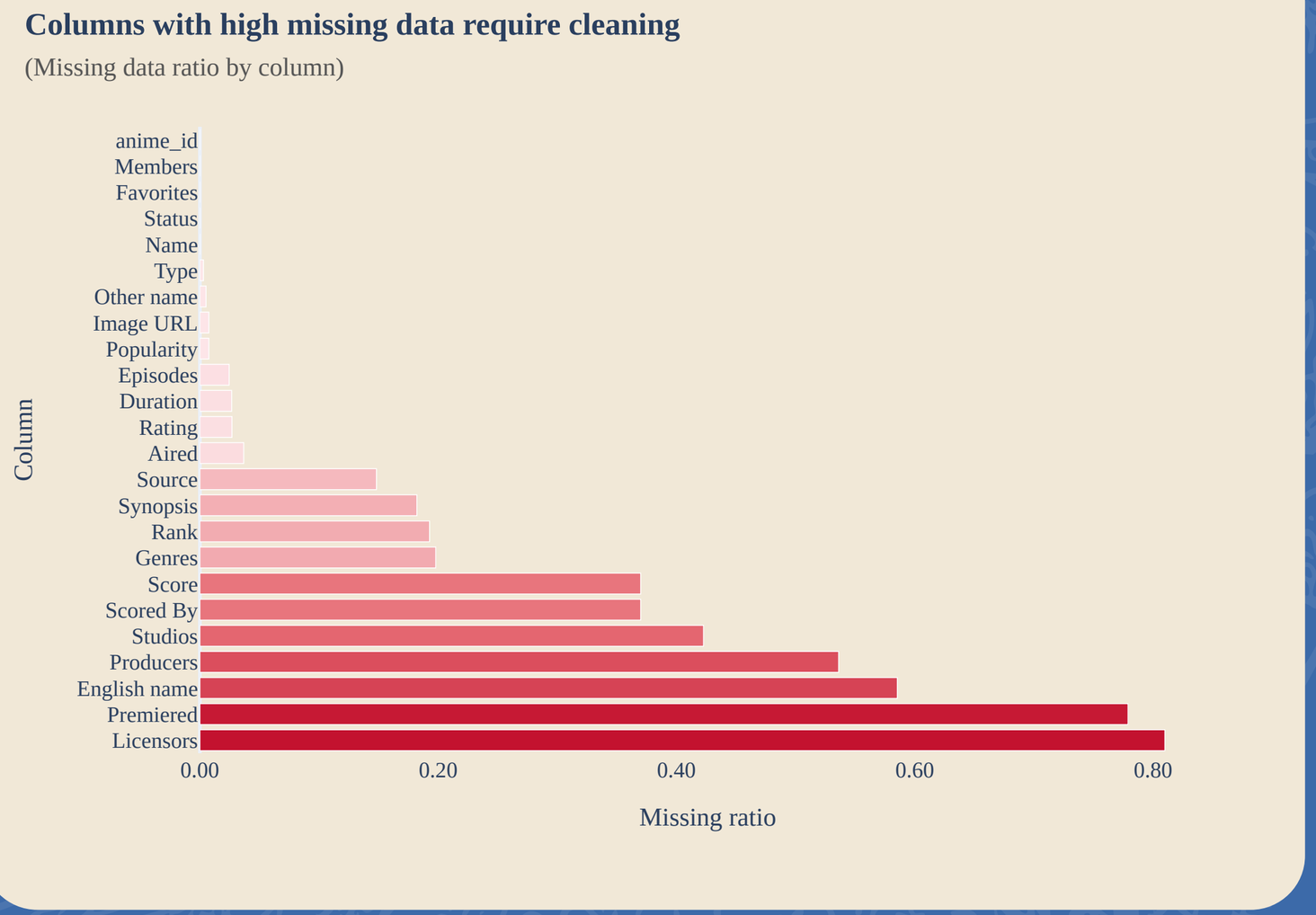
No description available for this anime.

Unknown Not available 0

https://cdn.myanimelist.net/img/sp/icon/apple-...

Score and Rank contradict each other

Rank = 190 Score = UNKNOWN



THE CONFLICT

Inconsistent, Fragmented, and Misclassified Metadata

Aired

15 formats

Inconsistent Metadata

Duration

12 min per ep

60 sec

1 hr

2 min per ep

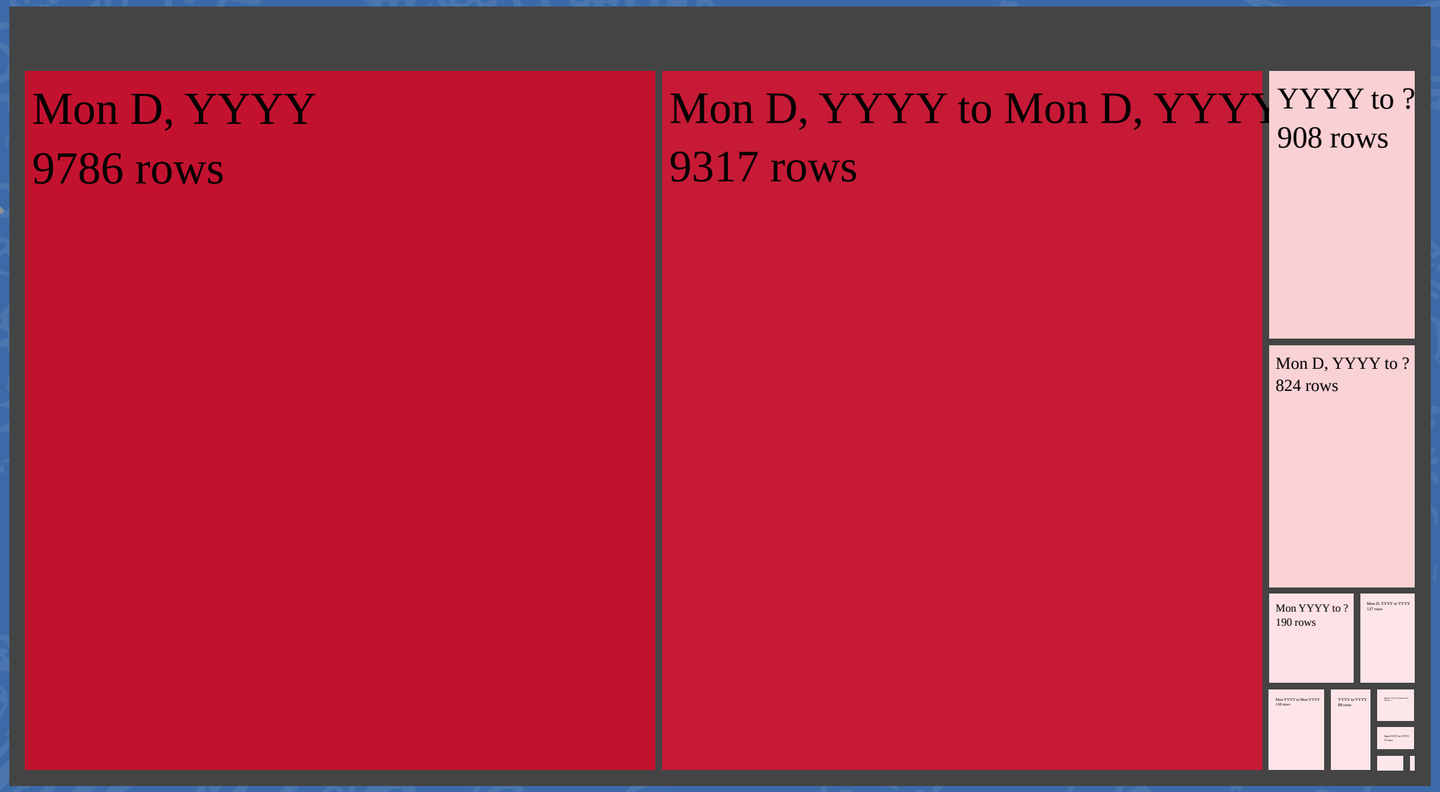
45 min

45 sec per ep

1 hr 3 min per ep

Studio

'Chongzhuo Animation'
vs
'Congzhuo Animation'



THE CONFLICT

Inconsistent, Fragmented, and Misclassified Metadata

Fragmented Metadata

- **Multi-label fields** explode into hundreds of rare or unique combinations.

Misclassified Metadata

Type, Episodes, and Duration contradict each other.

- Movies with multiple episodes
- TV entries with only a few minutes per episode
- Extreme episode counts

THE CONFLICT

Unreliable Distributions and Identity Ambiguities

Key engagement metrics (Members, Favorites, Episodes) are **heavily right-skewed**, making raw distributions statistically unstable.

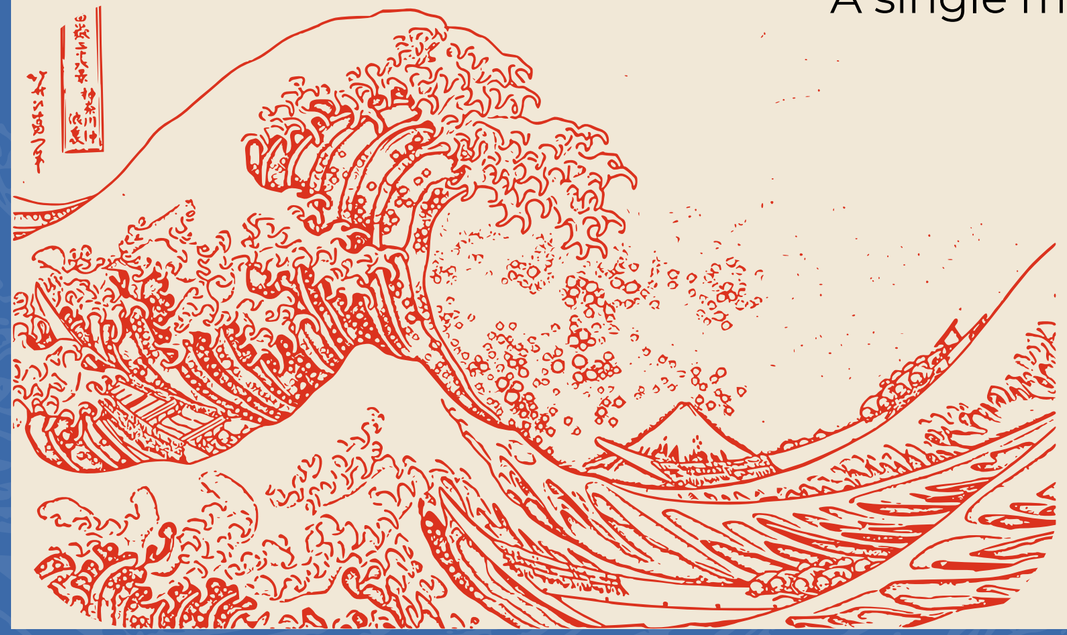
Multi-entity fields contain **ghost titles with zero engagement and duplicate names** masked by inconsistent metadata.

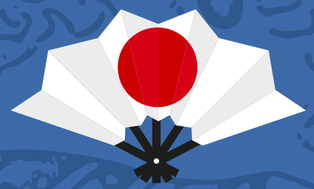
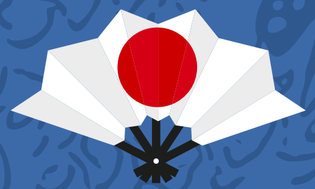
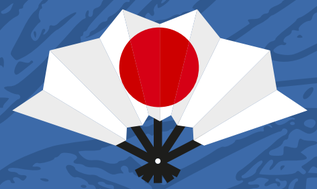
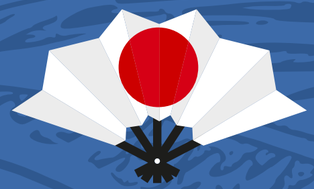
BIG IDEA

書き初め

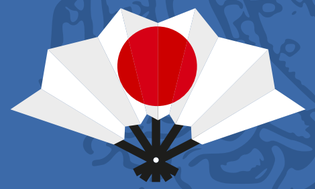
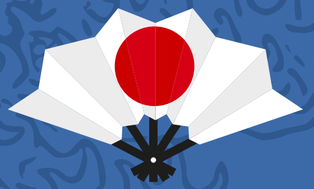
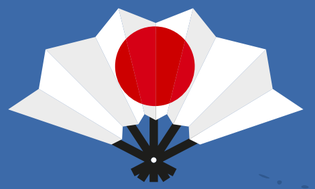
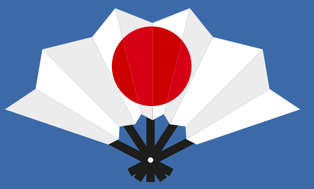
Only when cleaned and standardized does the data reveal what truly drives a high Anime Score.

A single message that guides the entire story.



3. CHAOS HANDLING

CHAOS HANDLING

Stage 1: Fixing Missingness
Expose true missing values, remove placeholders, restore critical fields.

Stage 2: Text and Structure Standardization
Normalize genres, studios, producers, and type labels.

Stage 3: Time Field Reconstruction
Parse Duration and Aired into consistent numerical formats.

Stage 4: Conflict Resolution
Correct misclassified Types, Episodes, and Durations.

Stage 5: Outlier and Distribution Repair
Identify extreme values and stabilize skewed metrics.

ACT 2

THE COMPLICATION & DISCOVERY

ACT 02

COMPARATIVE ANALYSIS

From Chaos to Clarity: A Journey of Discovery on Clean Data

- **Target Variable**
- **Market Factors**
- **Creative & Production Factors**
- **Release Strategy**

ACT 02

ANALYSIS STEPS

Issues Overview

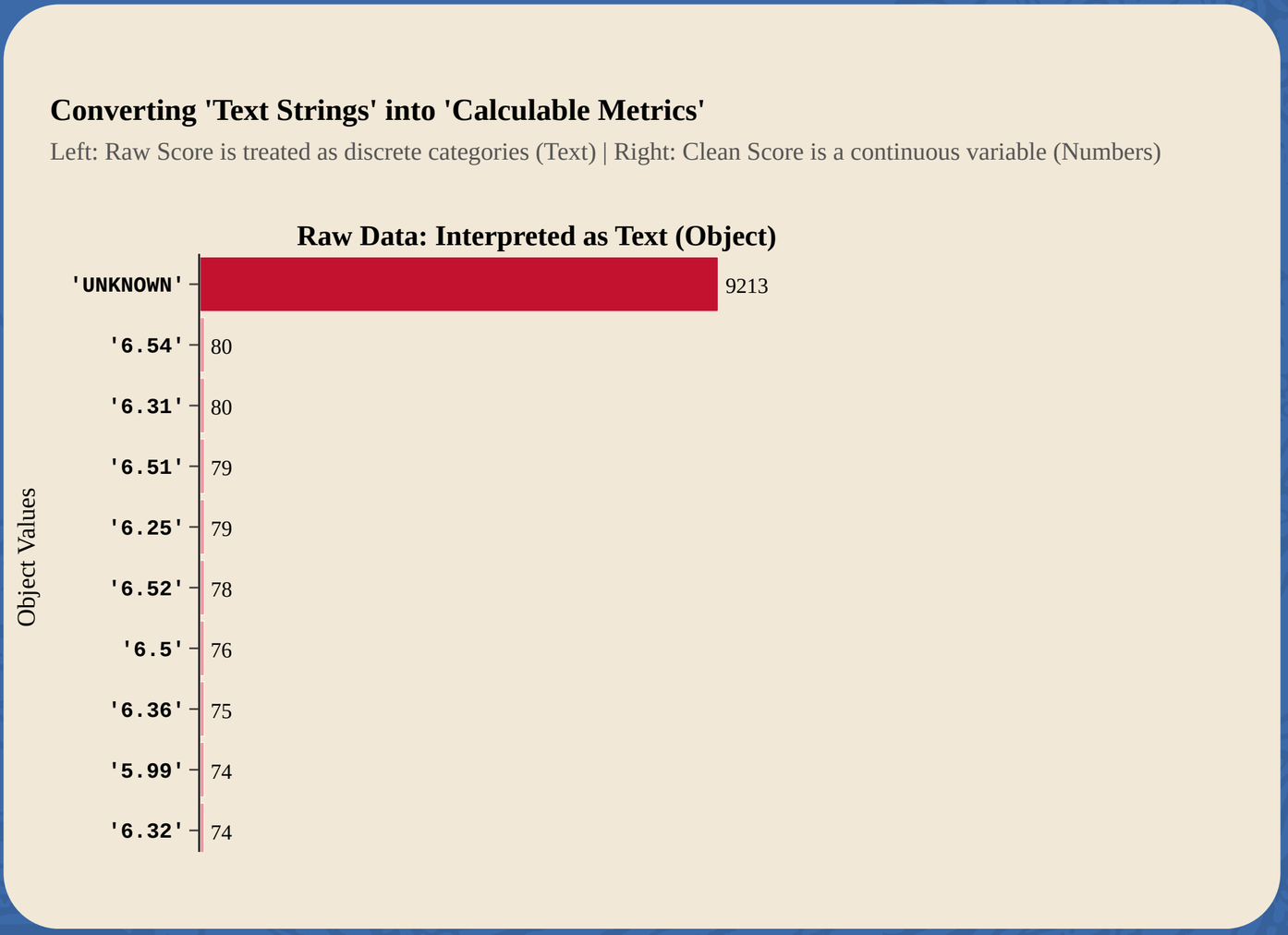
Solution

True Insight

Business Insight

1.TARGET VARIABLE-SCORE

1. SCORE - RAW DATA



Issues

- **37%** of the records is **unknown or missing values**
- **Inconsistencies in data types** (strings mixed with numbers)

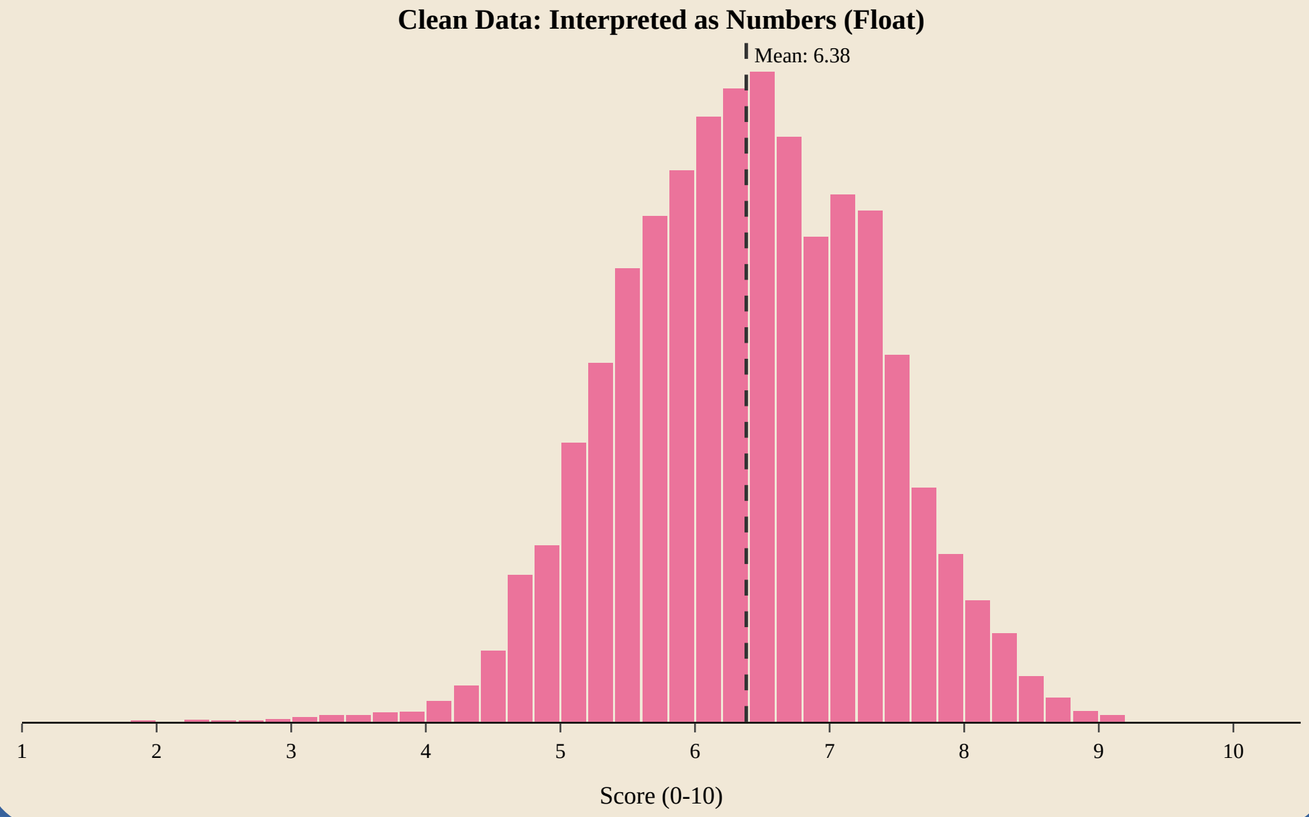
Consequence

- The **"Zero Trap"** made it impossible to calculate averages or find trends.

1. SCORE - CLEAN DATA

Converting 'Text Strings' into 'Calculable Metrics'

Left: Raw Score is treated as discrete categories (Text) | Right: Clean Score is a continuous variable (Numbers)



Solution

- **Standardization:**
  - Converting strings to numeric
  - Coercing unknown to NaN.
- **Filtration:** Drop all missing rows

Insight

- **Gaussian distribution**
- Average score is **6.38**

2. MARKET FACTORS

1

Type

Format types
(TV, Movie, etc.)

2

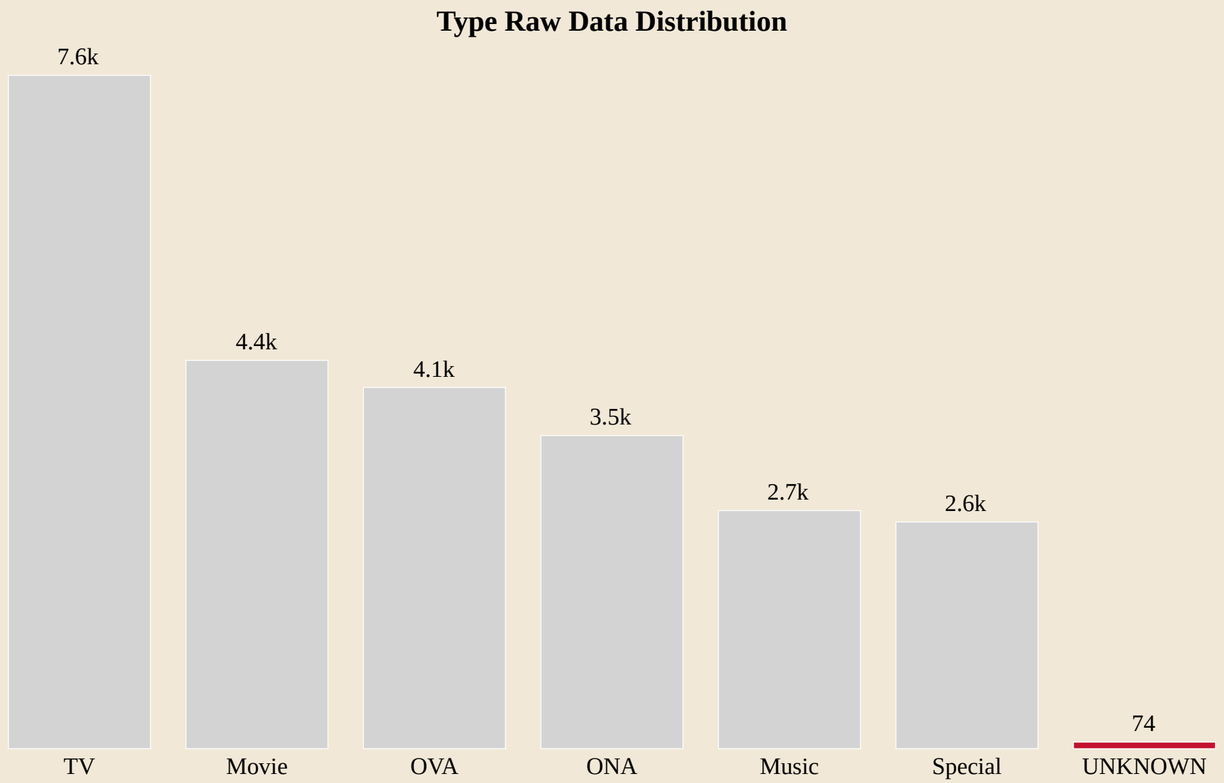
Source

Original source materials
(Manga, Novel, etc.)

2.1. TYPES - RAW DATA

Removing 'Unknown' Values Reveals True Type Distribution

(Top 10 Most Common Anime Types)



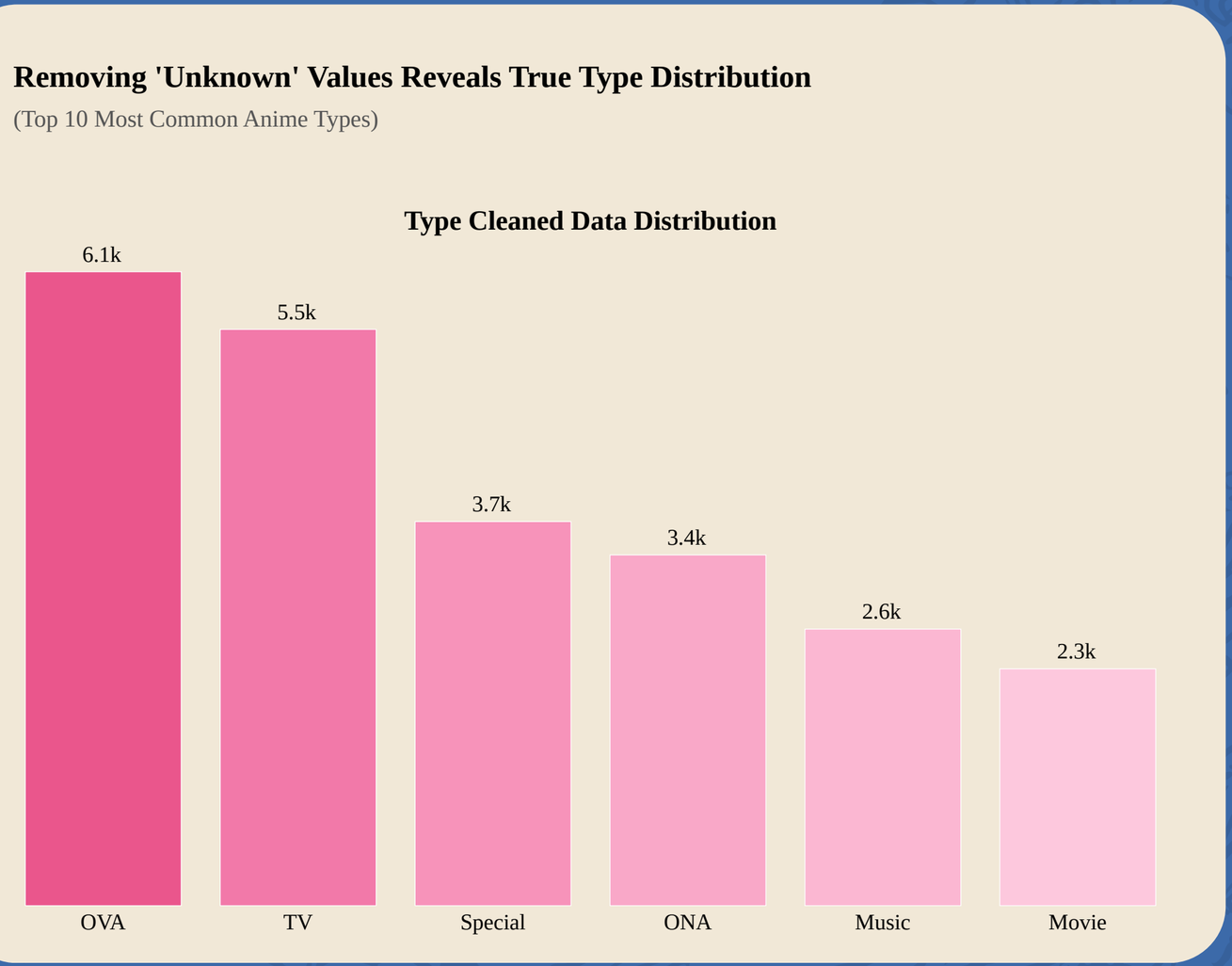
Issues

- **Misclassification: Short clips (< 40 mins)** were wrongly labeled as **"Movies"**
- "**Unknown"** text values distorted the count

Consequence

- Some Types volume are **not correctly counted**

2.1. TYPES - CLEAN DATA



Solution

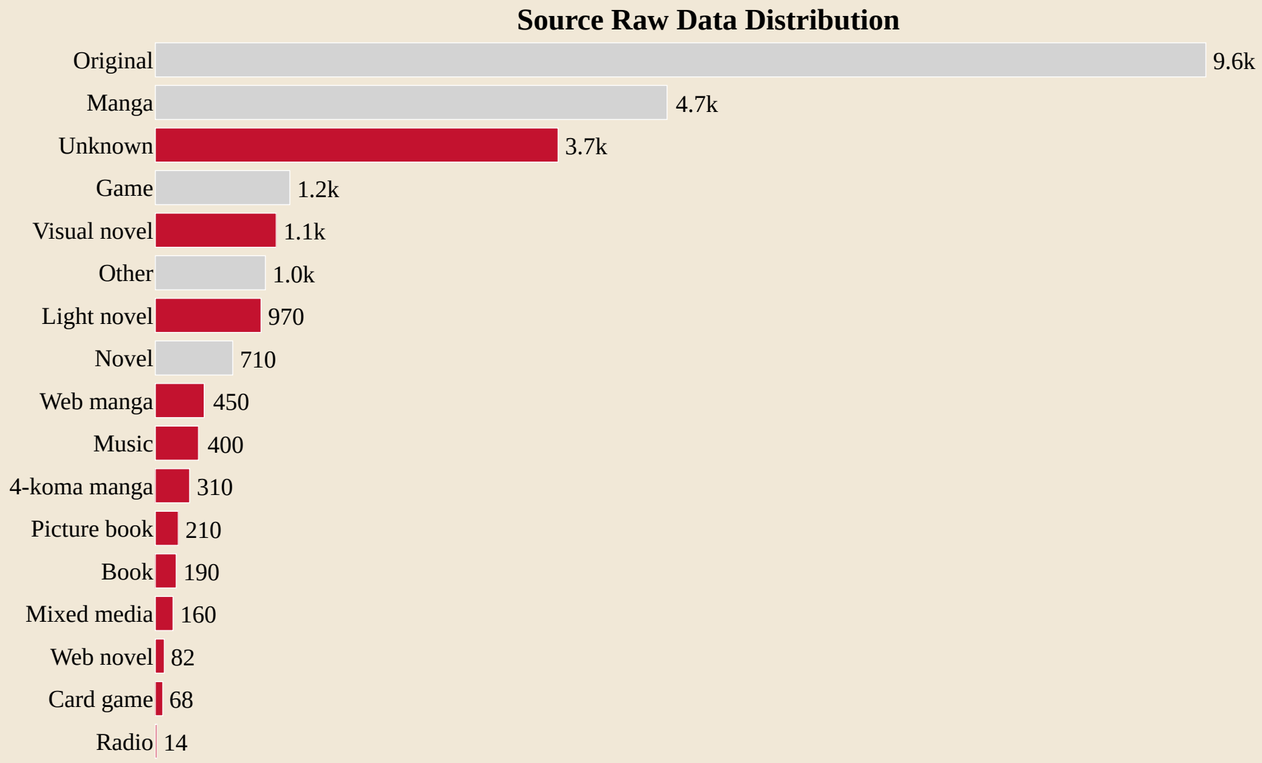
- **Reclassification:** If a "**Movie" < 40 minutes,** it is automatically changed to **"OVA"**
- **Standardization & Filtration:**
  - Convert the literal "Unknown" text into a standard NaN
  - Drop all missing rows

Insight

- **OVA #1 Types (6.k1)**
- All types have changed

2.2. SOURCES - RAW DATA

Consolidating Fragmented Sources Reveals True Market Structure



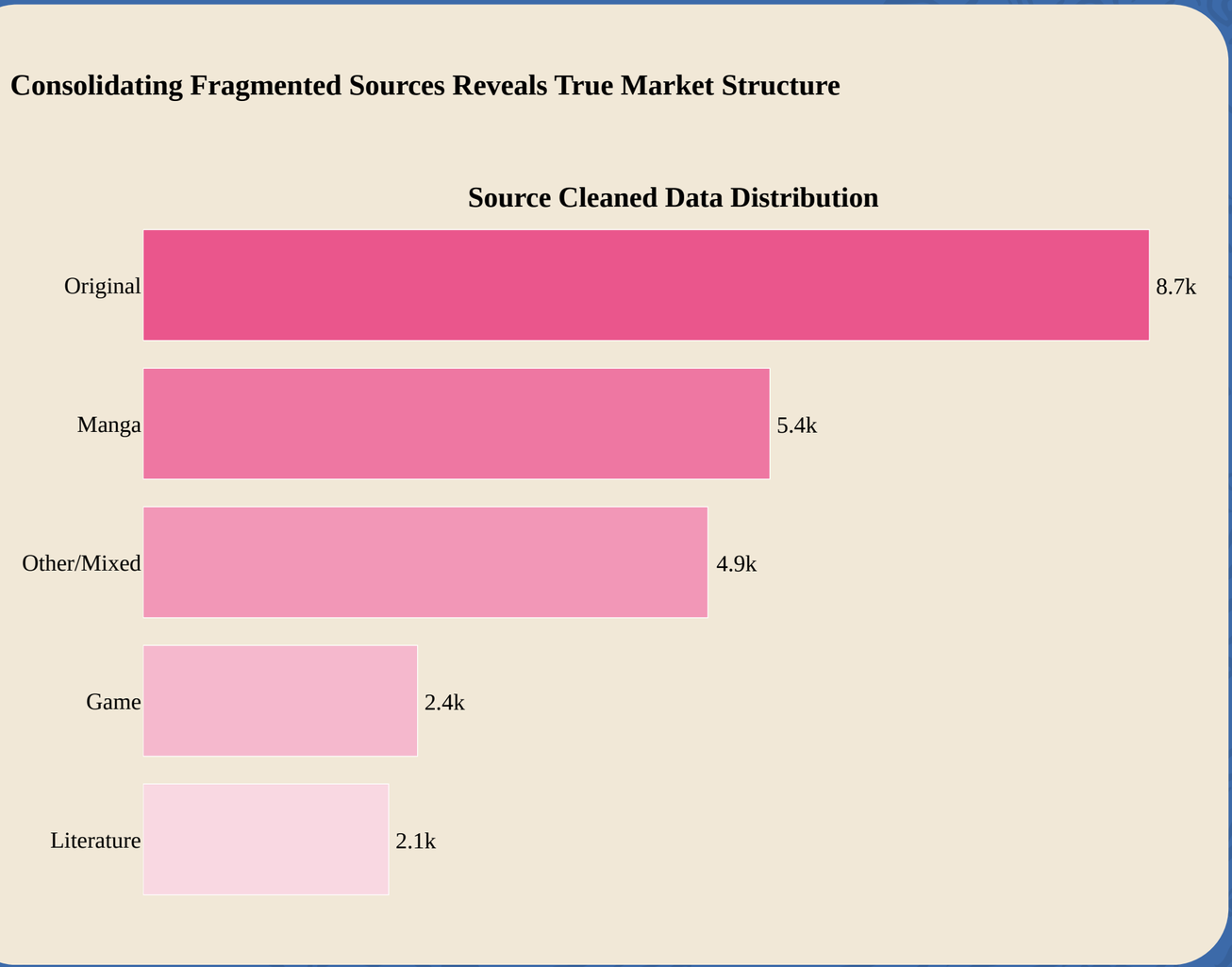
Issues

- **Issue:** The data was split into too **many small sub-categories.**
- "**Unknown"** text values distorted the count

Consequence

- Distorted the true market dominance.

2.2. SOURCES - CLEAN DATA



Solution

- **Consolidating** related types **into 3 Parent Groups** ((Manga, Literature, Game)
- Standardization & Filtration for **UNKNOWN**

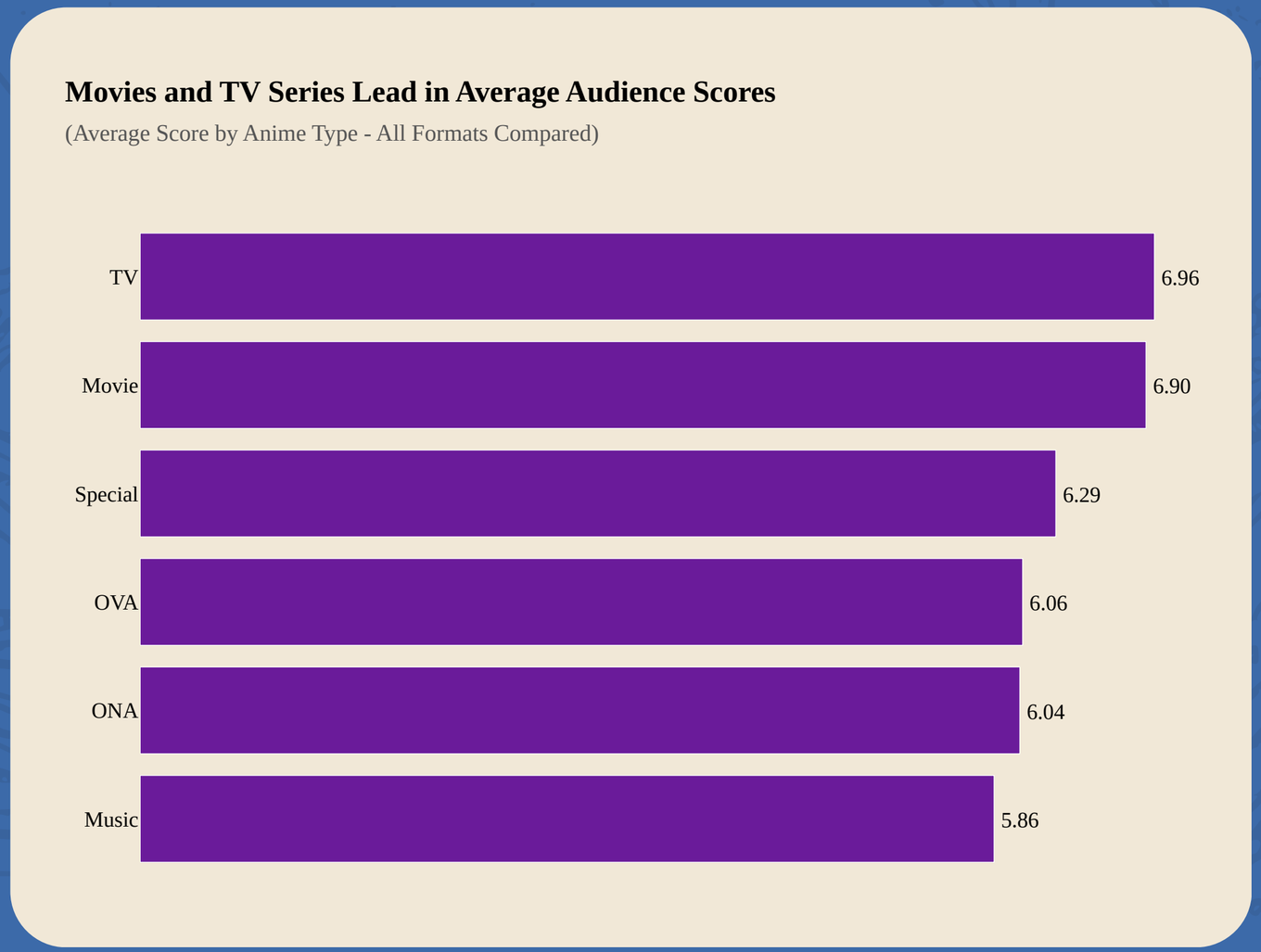
Insight

- **Original #1 Sources**
- Accurately compare the Market Share of all sources

2.3. BUSINESS INSIGHT

2.3.1. TYPE ANIME COMPARISON

Which formats maximize approval, and where is the Quality Ceiling?



Insight

1. The "Mainstream Prestige" Gap

- **"Format Tiering":** Primary (TV/Movies) vs Supplementary formats.
- **The Gap: TV/Movies ~7.00**, while others suffer a "Relevance Penalty" **(< 6.30)**.

2. Strategic Positioning: The Score Ceiling

- **Twin Pillars (TV 6.96 | Movie 6.90):** The only reliable paths to critical acclaim ("Main Events").
- **ONA/OVA Plateau (~6.05):** lack of broadcast prestige.
- **Music Floor (5.86):** Lowest → lack of narrative depth.

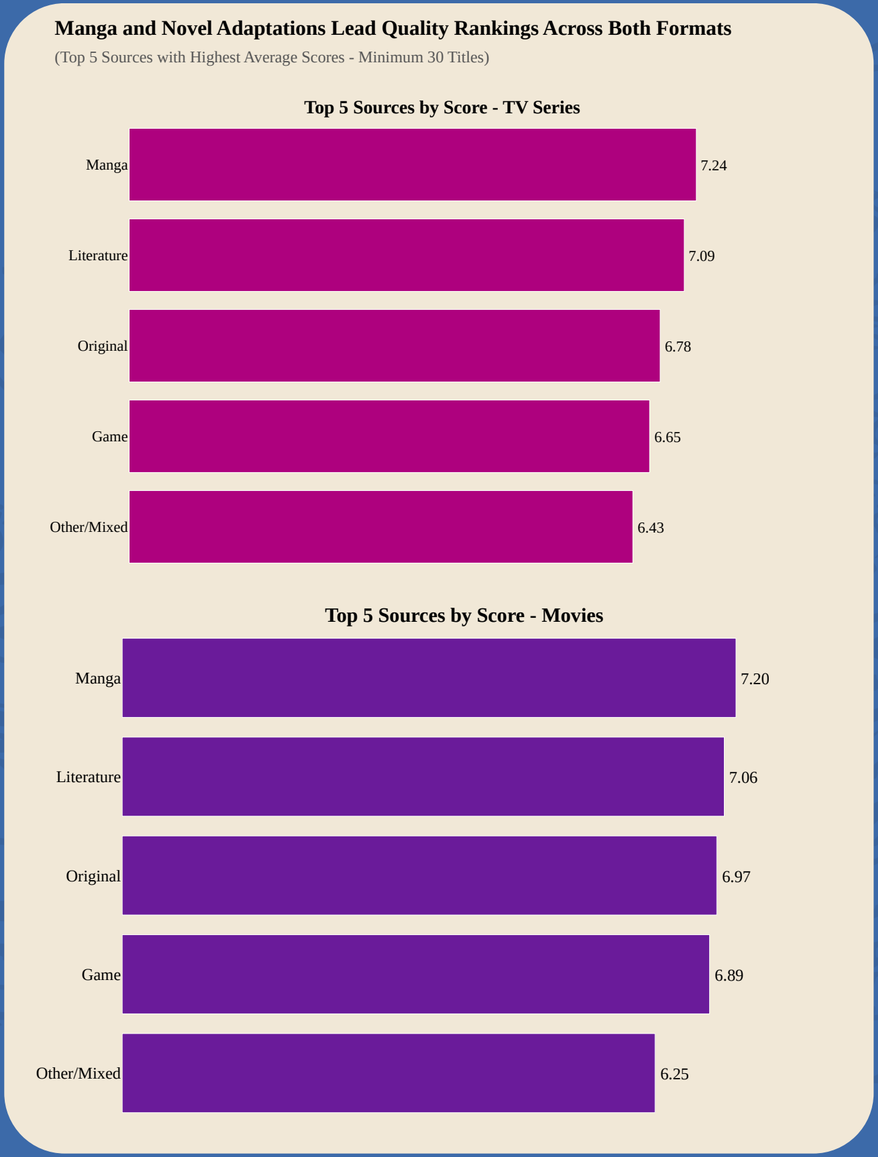
2.3.1. TYPE ANIME COMPARISON

Business Takeaways

- Concentrate resources on **TV & Movies** to break the **7.00 Score Barrier.**
- Treat **OVAs/ONAs** strictly as **fan engagement tools**, as they have a structural quality cap.

2.3.2. TOP SOURCE BY SCORE OF TV AND MOVIE

Which source materials consistently deliver the highest quality perception (Score)?



1. **The "Proven Narrative" Premium**

- **Adaptation Advantage: Manga & Literature** consistently outperform Originals and Games.
- **The Baseline: Existing storylines** offer a **"Quality Guarantee,"** maintaining scores above 7.00.

2. Strategic Positioning: Quality Drivers

- **Safety Net (Manga/Lit):** Undisputed leaders **(Manga TV: 7.24)**. Deep lore minimizes narrative failure
- **Originals:** Struggle in TV but thrive as **Movies (6.97).** Cinematic formats suit new stories better than series
- **Game : Consistently lag behind.** Converting "gameplay" to story remains a structural weakness

2.3.1. TYPE ANIME COMPARISON

Business Takeaways

- **Priority: Manga or Literature** maximizes score potential.
- **Strategy:** If **Original IP**, prioritize the **Movie** format over TV Series.
- **Caution: Game adaptations** face a historical quality cap; they require extra **investment** in **screenwriting.**

3. CREATIVE & PRODUCTION FACTORS

1

Genres

(Comedy, Fantasy, Romance, etc.)

2

Producers

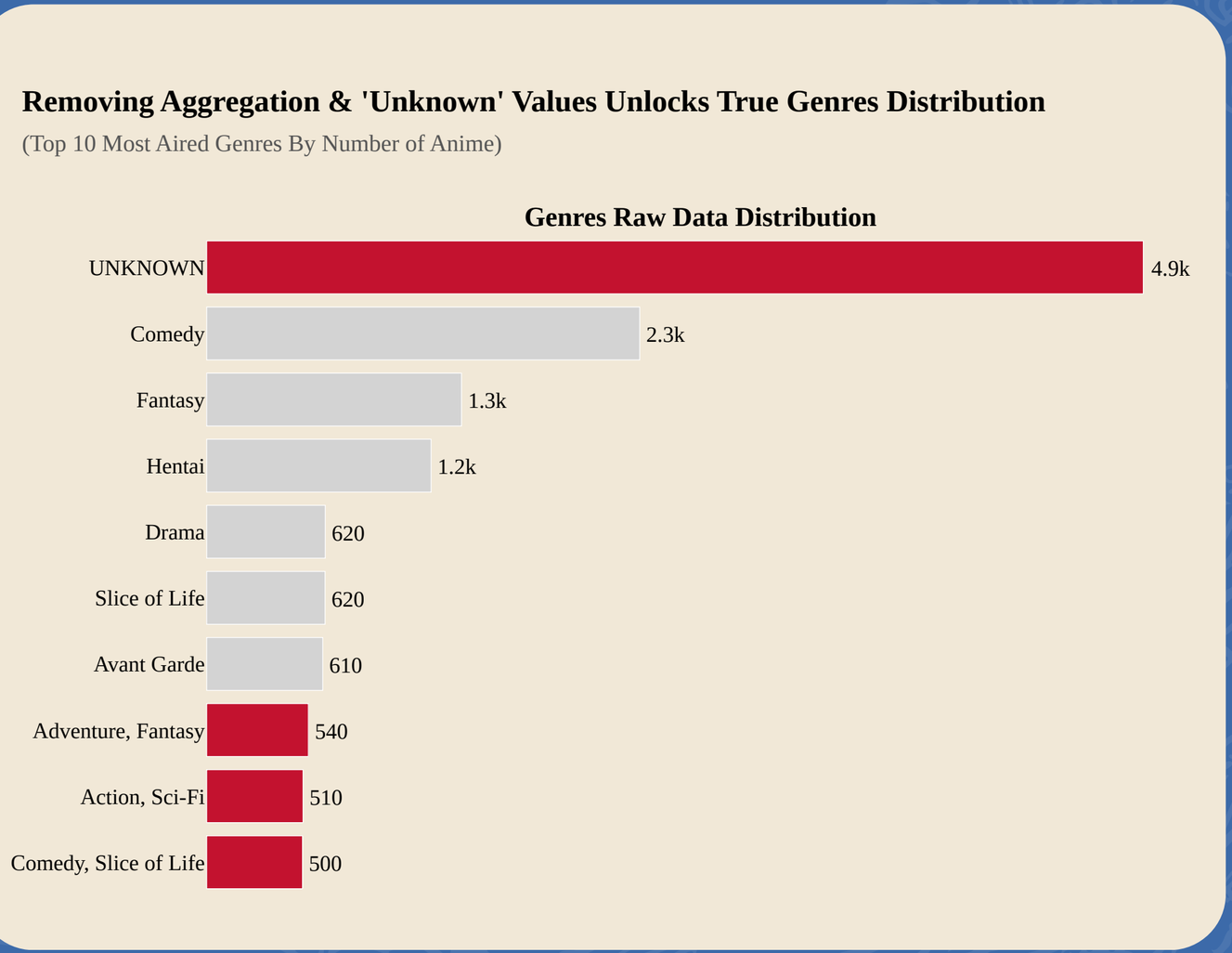
Produce or fund the anime

3

Studios

Directly responsible for visual content

3.1. GENRES - RAW DATA



Issues

- **UNKNOWN dominance**
- **Aggregated lists:** Genres were locked inside stringified lists **(e.g., ['Action', 'Comedy'])**

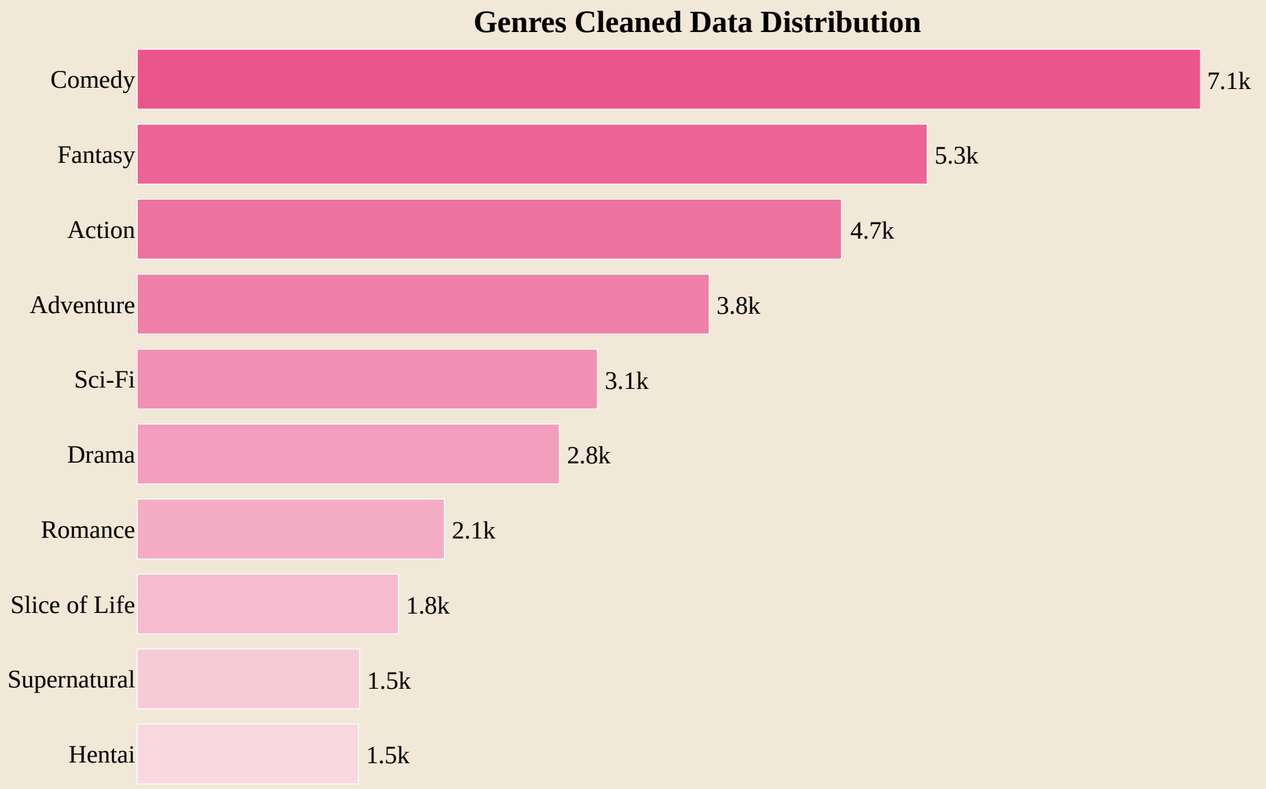
Consequence

- **Masking the actual market structure**
- Individual genres **didn't get proper credit**

3.1. GENRES - CLEAN DATA

Removing Aggregation & 'Unknown' Values Unlocks True Genres Distribution

(Top 10 Most Aired Genres By Number of Anime)



Solution

- **Regex: S**tripped brackets **[ ] and** quotes **' '**
- **Explode:** Split lists into **individual rows**
- **Trim:** Removed extra **whitespace**
- **Standardization & Filtration** for **UNKNOWN**

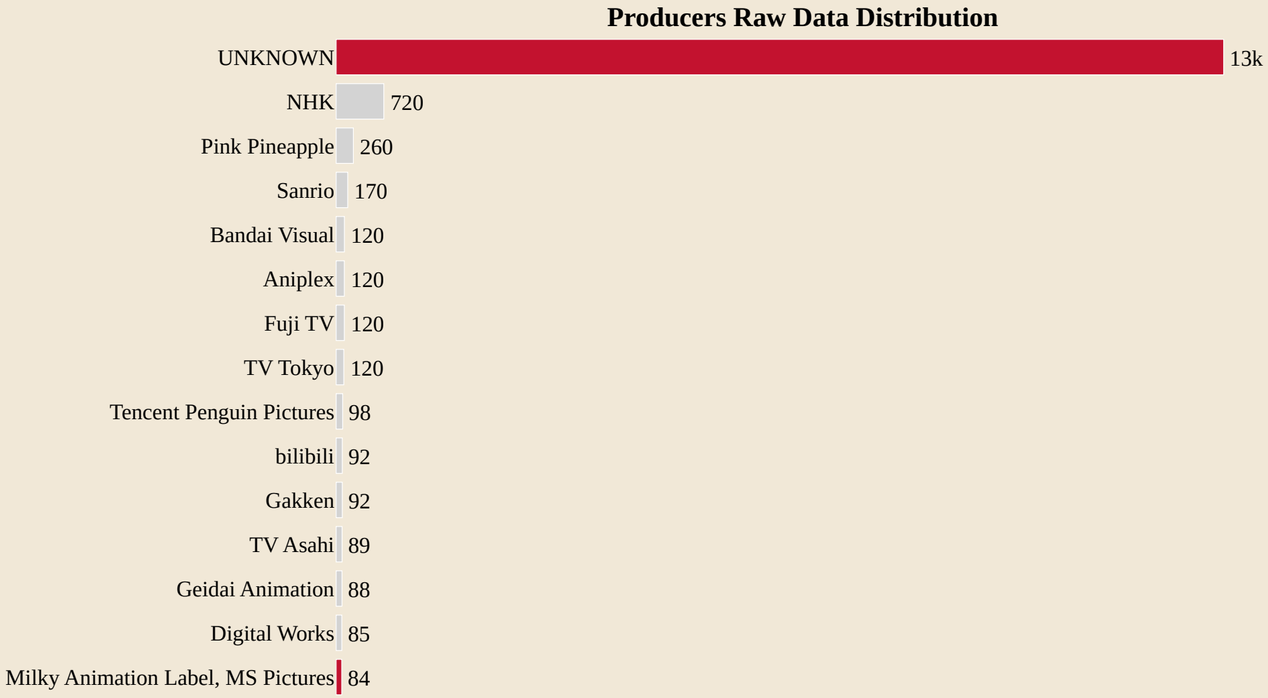
Insight

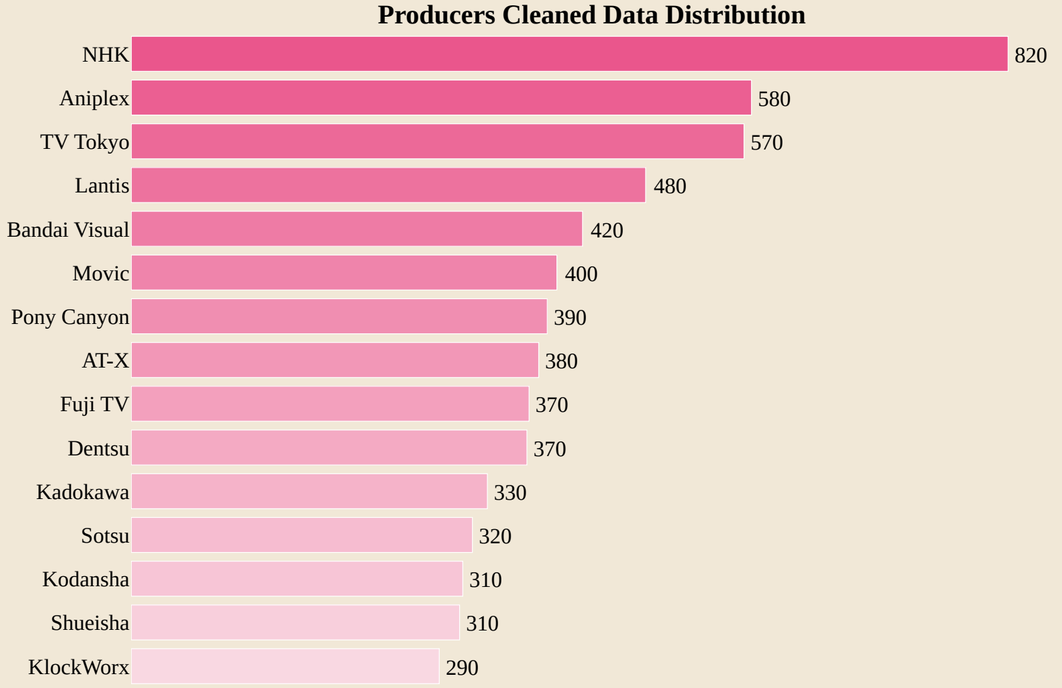
- **Comedy #1 Genre (7.1k)**
- **Accurately record** the total volume of anime by each genre

3.2. PRODUCERS - RAW & CLEAN DATA

Parsing Producers Data Reveals True Market Leaders

(Top 15 Most Active Producers)



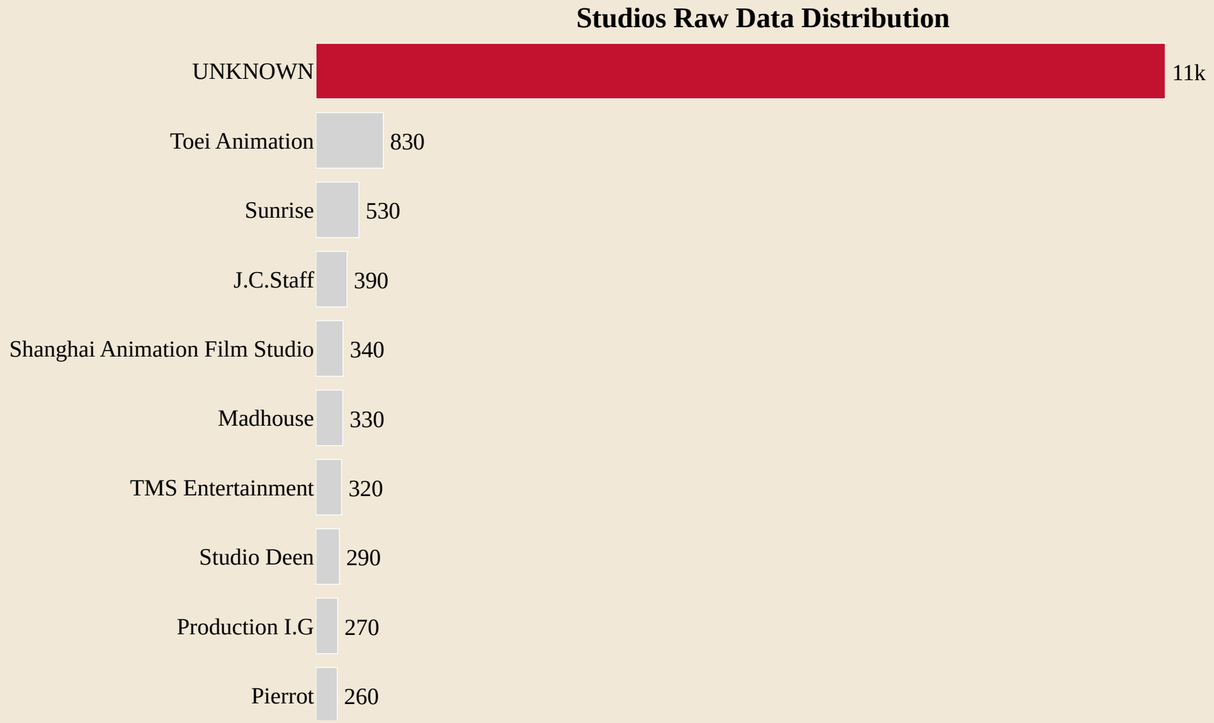


- **NHK #1 Producers (820)**
- All producers are **correctly attributed** to the number of invested projects

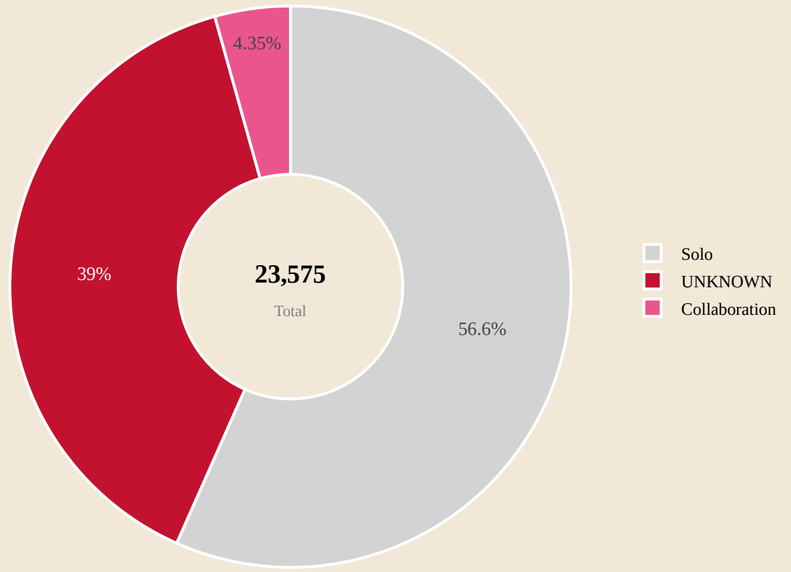
3.3. STUDIOS - RAW DATA

Parsing Studio Data Reveals True Studios Market Leaders

(Top 10 Most Active Studios)



Raw Data: UNKNOWN Dominance and Aggregation In Studio's Proportion

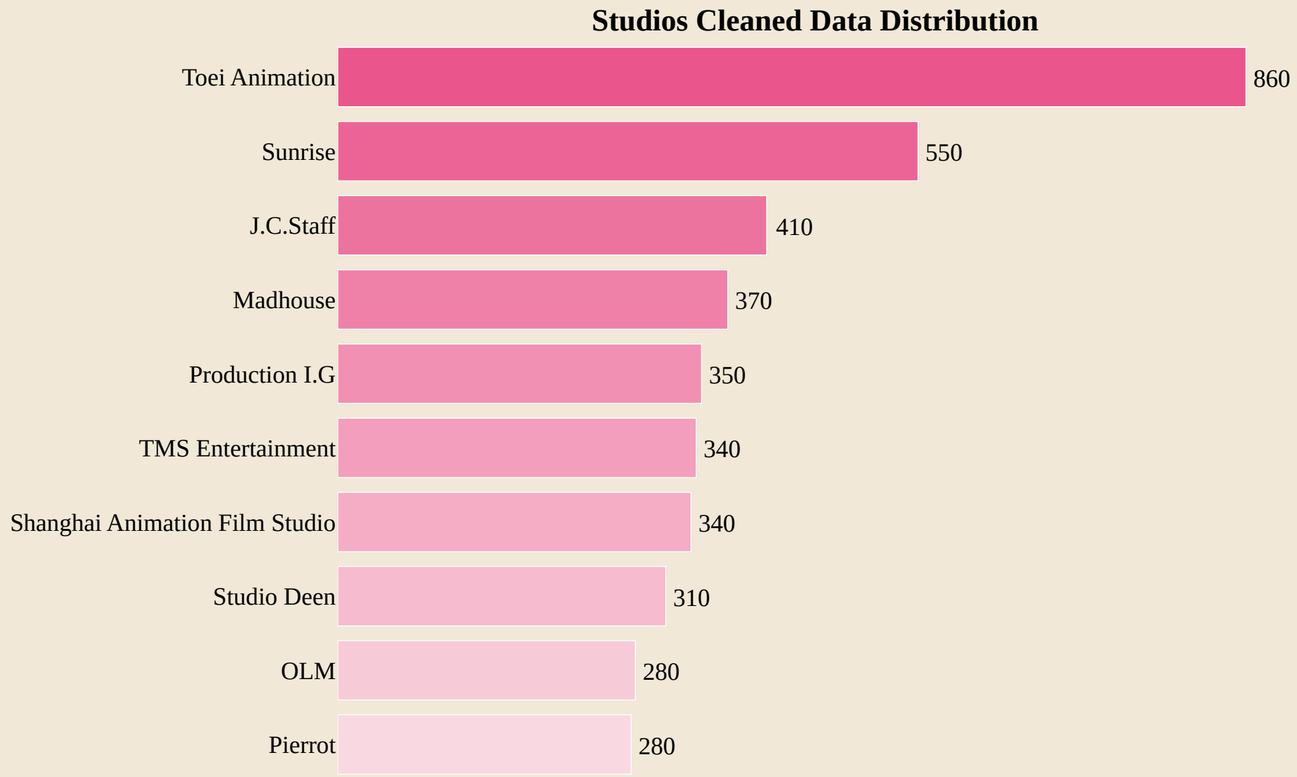


- **UNKNOWN and Solo studios** account for a large portion
- Appearance of **Collaborations**

3.3. STUDIOS - CLEAN DATA

Parsing Studio Data Reveals True Studios Market Leaders

(Top 10 Most Active Studios)



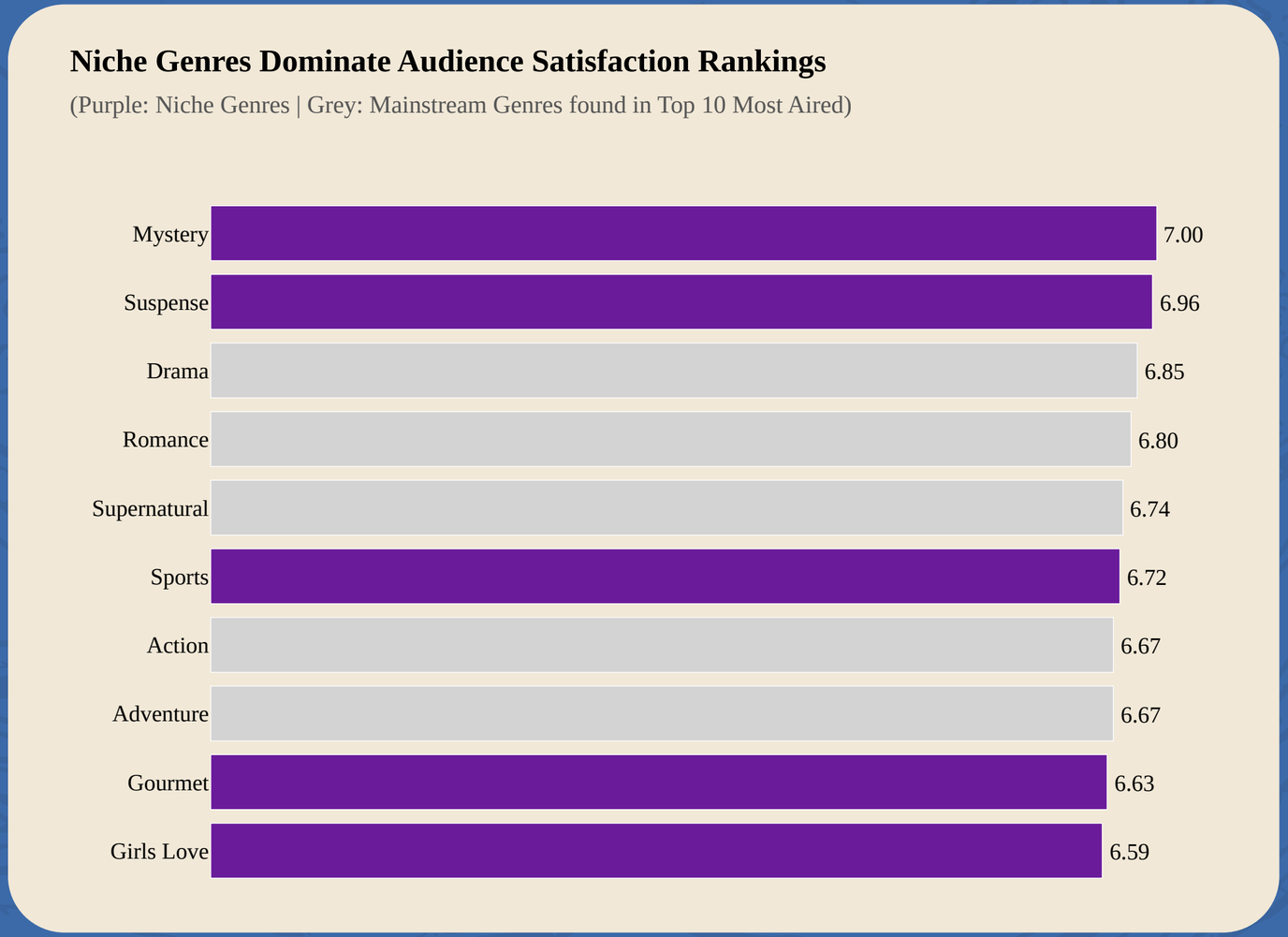
True insight

- **Toei Animation #1 Studios (860)**
- All studios are **correctly attributed** to the number of invested projects

3.4. BUSINESS INSIGHT

3.4.1. GENRES VS SCORE ANALYSIS

What are the emerging genres with high assessment?



Insight

- **Niche Wins at the Top: Mystery** and **Suspense r**eceive the highest scores. **Specialized content** drives the strongest fan engagement
- **The Safe Zone: Drama and Romance** are reliable choices. They consistently deliver high scores without high risk.

3.4.1. GENRES VS SCORE ANALYSIS

Business Takeaways

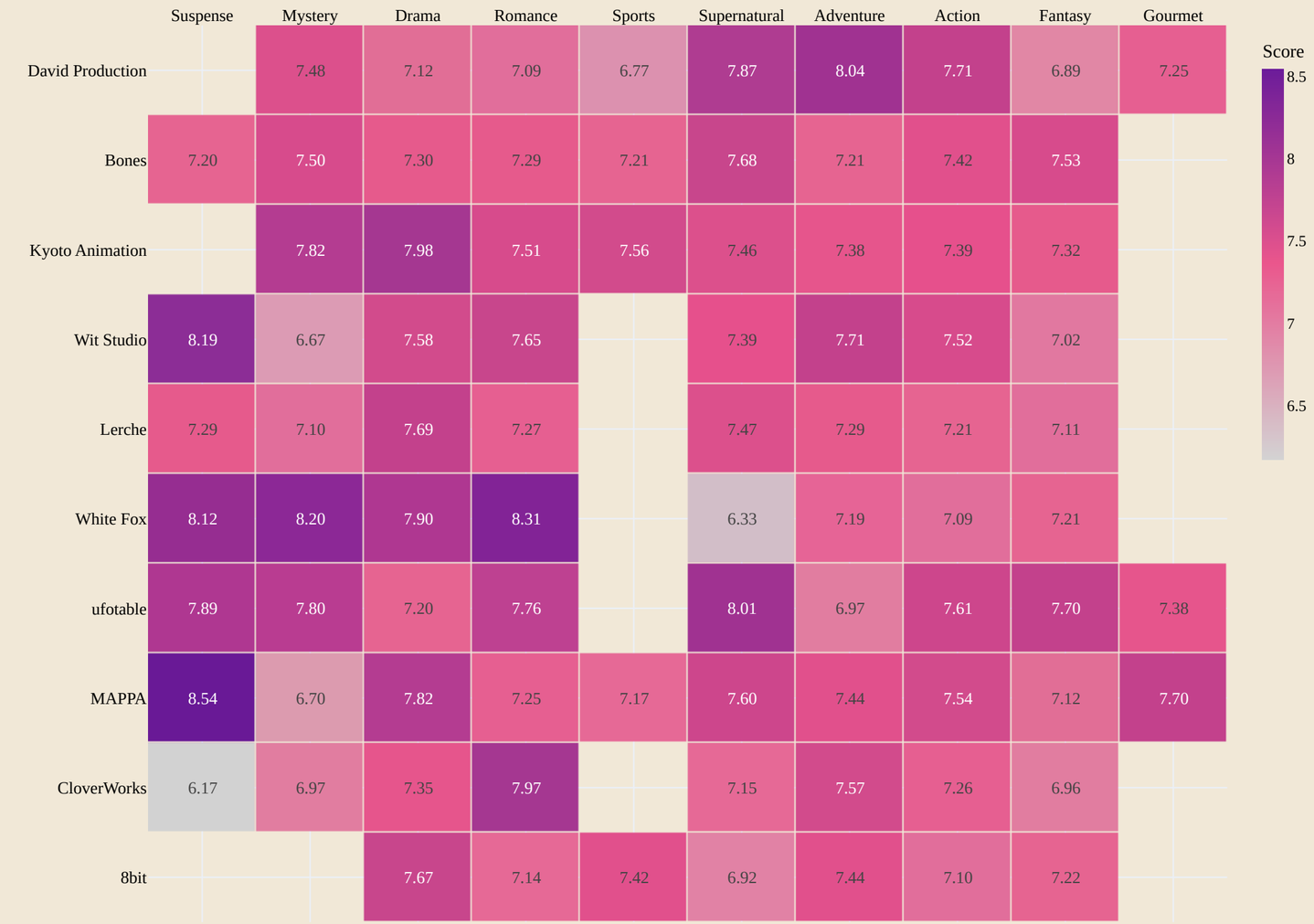
- **Max Score Strategy: Prioritize Mystery or Suspense** to chase the highest possible rating.
- **Safe Bet Strategy:** Invest in **Drama or Romance** for a balance of broad appeal and quality.
- **Hybrid Strategy: Create "Super-Niche" titles.** Mix deep Mystery elements with Mainstream themes to get the best of both worlds.

3.4.2. STUDIOS VS GENRES

Which Genres do top Studios consistently excel in?

Elite Quality Intersection: Top Studios vs. Top Genres

(Performance Matrix of the 10 Highest-Rated Studios across the 10 Highest-Rated Genres)



1. Absolute Peak Performance: High-Tension Genres

- **Scores > 8** are concentrated in **Suspense and Mystery.**
- **MAPPA -** highest score **(8.54** in Suspense)

2. White Fox: The Versatility King

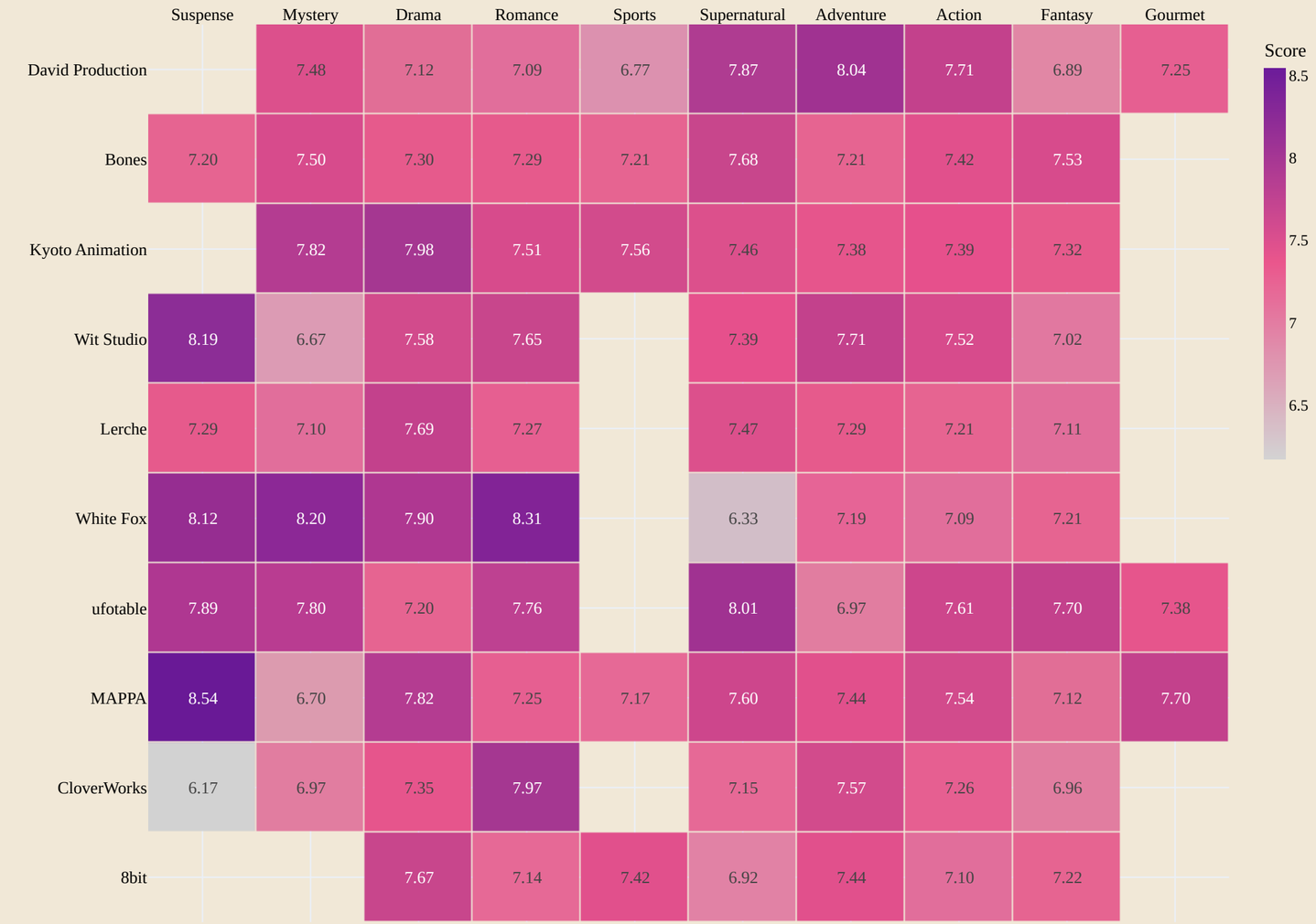
- White Fox is the only studio to break **8.00 i**n two opposing genres: **Romance (8.31)** and **Mystery (8.20).**

3.4.2. STUDIOS VS GENRES

Which Genres do top Studios consistently excel in?

Elite Quality Intersection: Top Studios vs. Top Genres

(Performance Matrix of the 10 Highest-Rated Studios across the 10 Highest-Rated Genres)



3. Hidden Risks: Studio-Genre Mismatches

- The Trap: Even elites have "danger zones." **CloverWorks** hits the matrix low in **Suspense (6.17).**

4. Niche Genres

- The Gap: Very **few** studios achieve high scores **(>7.50)** in **Sports or Gourmet.**

3.4.2. STUDIOS VS GENRES

Business Takeaways

1. Strategy 1: Target the "Triple-A Gold."

- **Focus on the top 3 proven combinations for maximum scores:**
  - MAPPA + Suspense (8.54)
  - White Fox + Romance (8.31
  - White Fox + Mystery (8.20)

2. Strategy 2: The "Smart Crossover" (Niche Expansion)

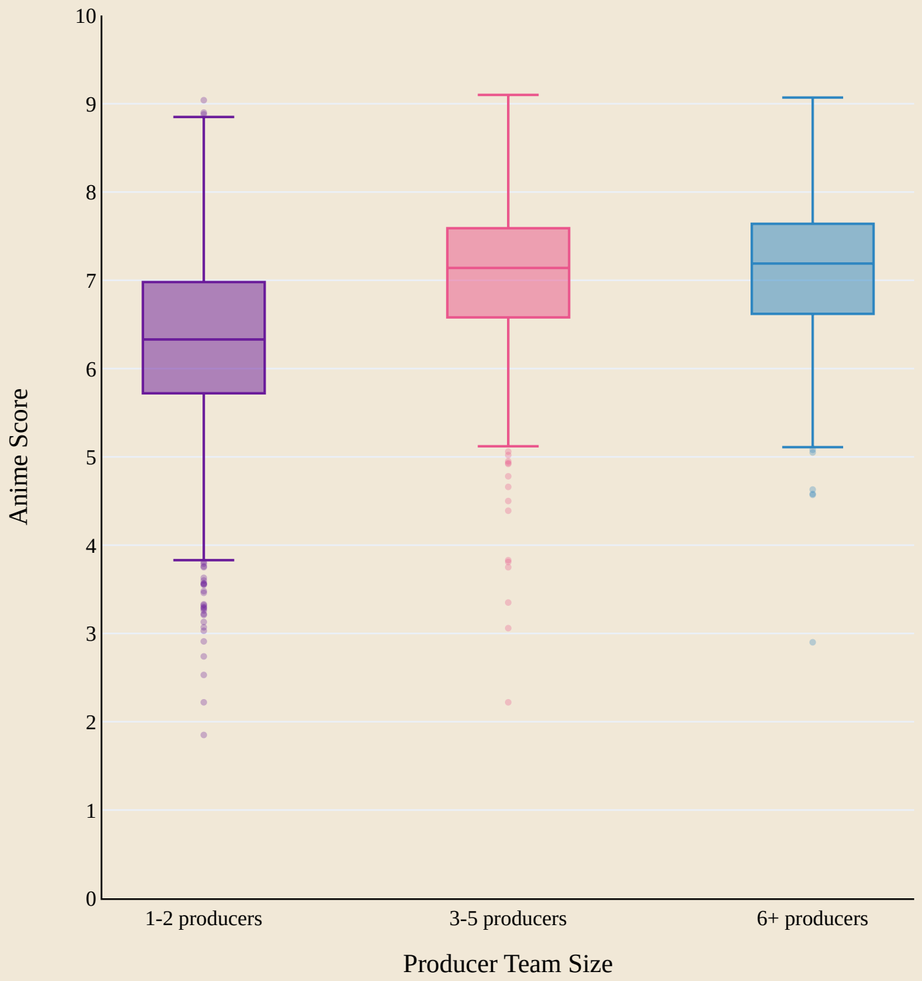
- **Leverage core technical skills to dominate less competitive genres:**
  - **Action → Sports:** High-motion animation skills transfer perfectly to Sports. (Wit Studio & Bones)
  - **Drama → Gourmet:** Strong attention to detail and emotion fits Gourmet storytelling (KyoAni & WhiteFox)

3. Strategy 3: The "Specialist" Rule

- **To ensure stability and quality, only partner with established leaders:**
  - Adventure: David Production.
  - Supernatural: Ufotable (for high-fidelity spectacle).
  - Drama: Kyoto Animation (The safest bet for consistent quality).

3.4.3. PRODUCERS COMMITTEE

Does Committee Size Impact the Score?



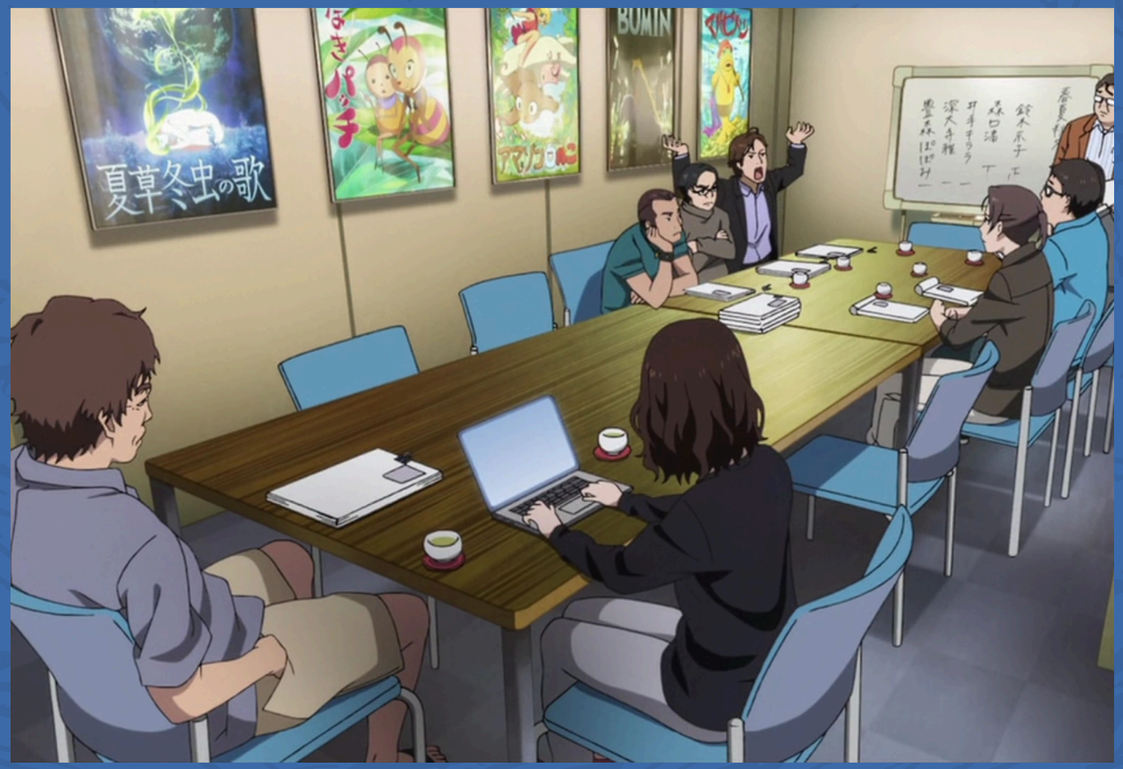
1. Insight: Larger Teams Deliver Better Results

- **The Performance Gap: 3-5 Producers (7.19)** significantly outperform **1-2 Producers (6.33).**
- **The Driver:** This **+0.81** confirms that **pooled resources and stricter quality control** lead to **a superior product.**

2. Insight: Risk Mitigation & Quality Floor

- **The "Solo" Risk: Small teams (1-2)** show high volatility, with "disaster" **outliers** plummeting to the **2.0–3.0 range.**
- **The Safety Net:** Committees of **3+ Producers** raise the "Quality Floor" to **~5.1.** Collaboration effectively protects the investment from catastrophic failure.

3.4.3. PRODUCERS COMMITTEE



Business Takeaways

- **Optimal Strategy:** Form a Production Committee of **3-5 Partners.**
- **Rationale: T**his structure is the "Sweet Spot," **balancing robust financing with management efficiency.**

4. RELEASE STRATEGY

1

Aired Date

2

Episodes

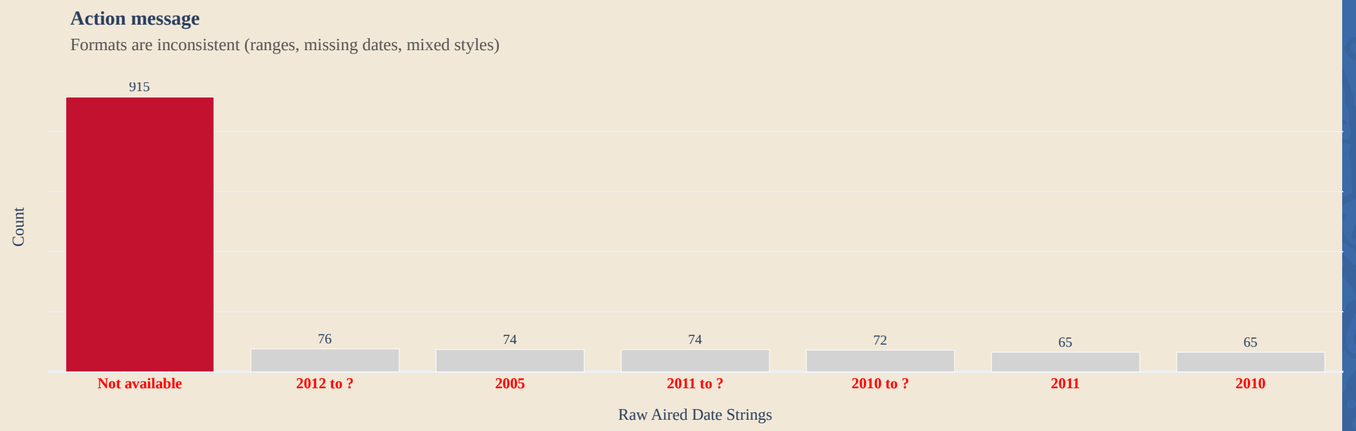
3

Duration

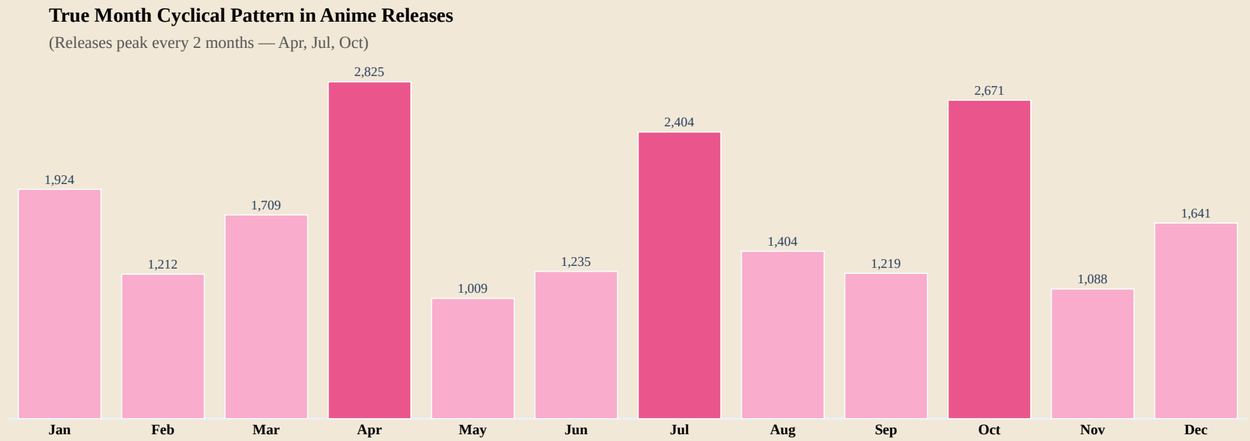
4.1. VISUAL EVIDENCE

AIRED

Raw data



Cleaned data



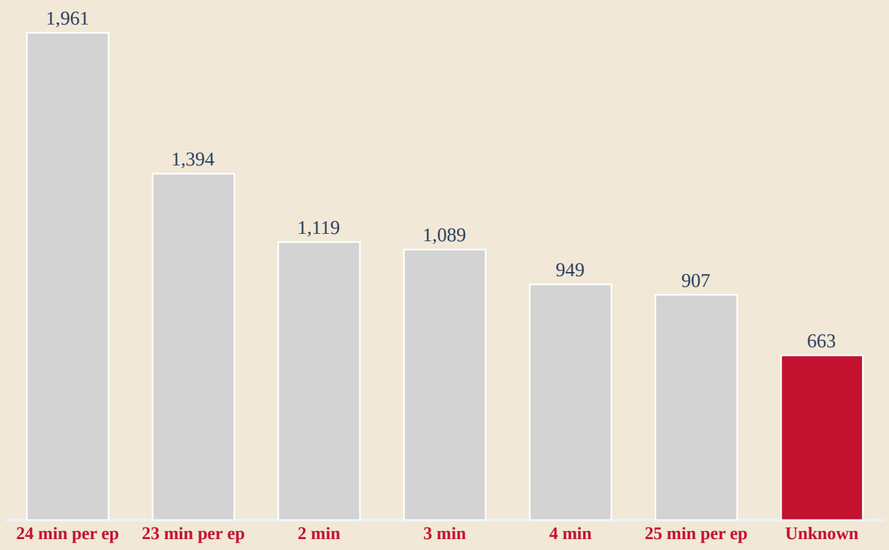
- **The Misleading View**: Raw string-based dates scrambled the timeline, hiding all seasonal trends.
- **The Truth After Cleaning**: After cleaning, a clear release pattern emerges with strong peaks in Jan, Mar, and Oct.
- **Strategic Value**: Aired Month becomes a reliable feature for identifying optimal release windows and industry seasonality.

4.1. VISUAL EVIDENCE

DURATION

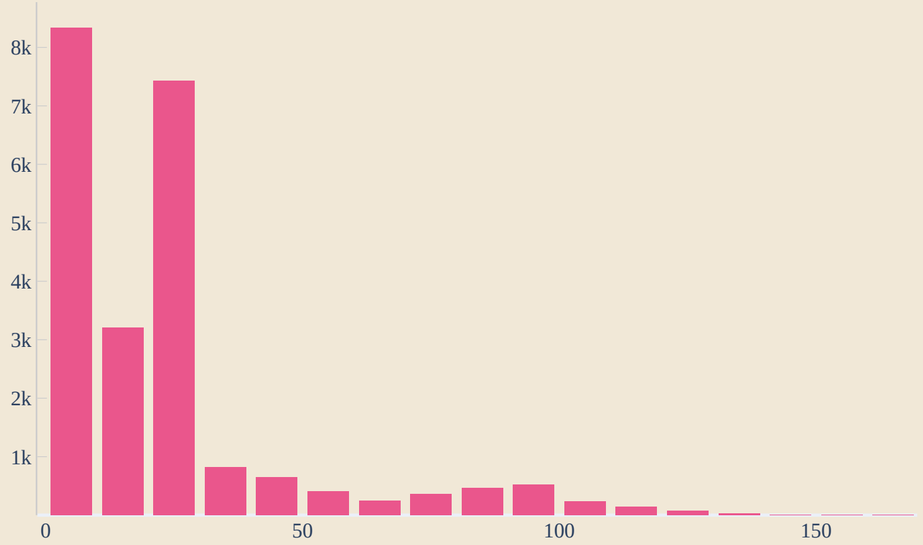
Duration Distribution is Distorted (Raw Data)

(Top 7 Duration Label Frequencies)



True Duration Distribution is Right Skew

(Cleaned Numeric Duration Distribution)



The Misleading View: Raw string-based dates scrambled the timeline, hiding all seasonal trends.

The Truth After Cleaning: the actual distribution is heavily right-skewed

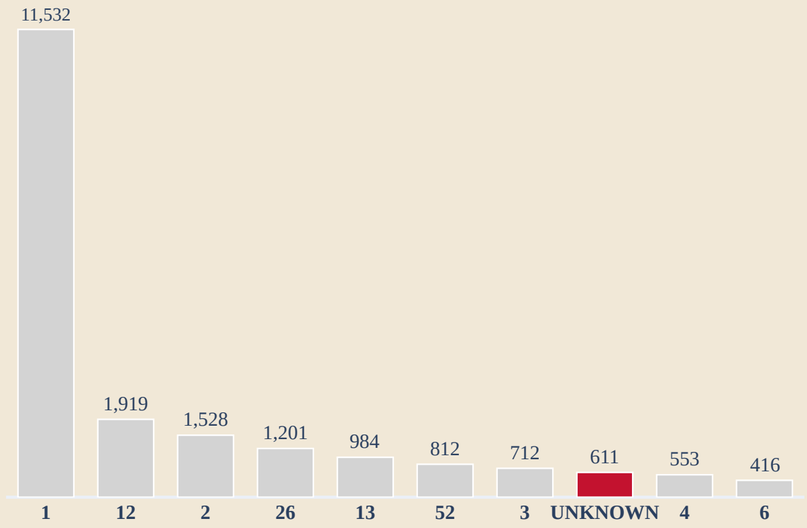
Strategic Value: Analyzing optimal episode length strategies separately by anime type

4.1. VISUAL EVIDENCE

EPISODES

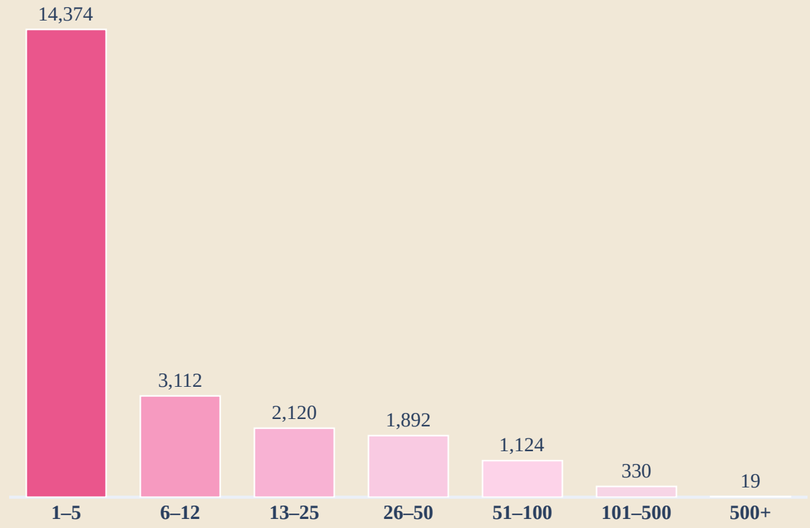
Episode Distribution is Distorted (Raw Data)

(Top 10 Episode Label Frequencies)



Mini-Series Dominate Anime (Cleaned & Binned)

(Episode Frequencies Across Groups)



The Misleading View

Raw episode strings and “UNKNOWN” labels

The Truth After Cleaning

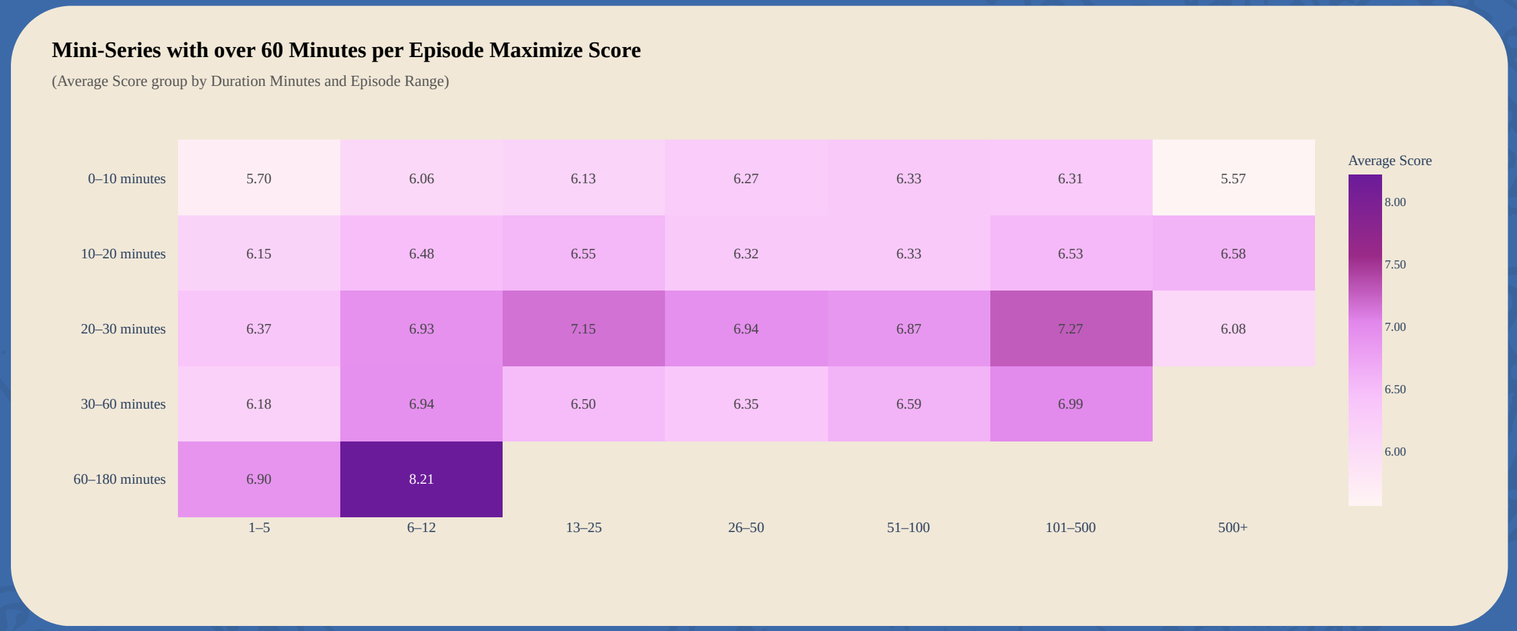
Mini-series contain the highest number of titles

Strategic Value

Analyzing length vs performance.

4.2. BUSINESS INSIGHTS

WHICH COMBINATION OF DURATION MINUTES AND EPISODE RANGE MAXIMIZE THE SCORE ?



Insight

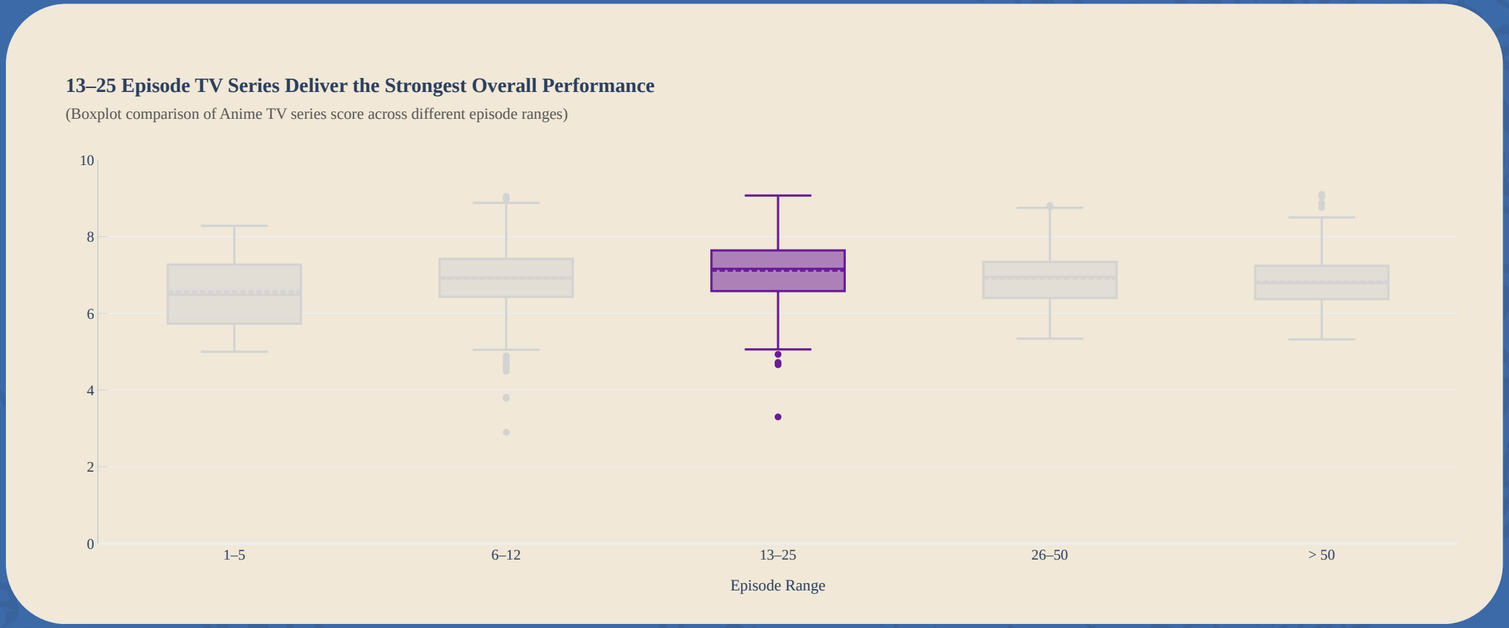
- Mini-series with long durations (60–180 min) and 6–12 episodes deliver the highest scores.
- Standard episodes (20–30 min) perform best when limited to 6–25 episodes.
- Ultra-short episodes (0–10 min) consistently score lowest.
- More episodes does not mean higher score; quality and pacing matter more than length.

Business Takeaways

- Prioritize premium limited series with longer runtimes.
- Avoid ultra-short formats as core productions.
- For standard-length episodes, keep series within 12–25 episodes for optimal reception.
- Invest in quality over quantity; pacing and duration drive audience satisfaction.

4.2. BUSINESS INSIGHTS

TV ANIME EPISODE STRATEGY: HOW LENGTH AFFECTS SCORE CONSISTENCY AND RISK



Insight

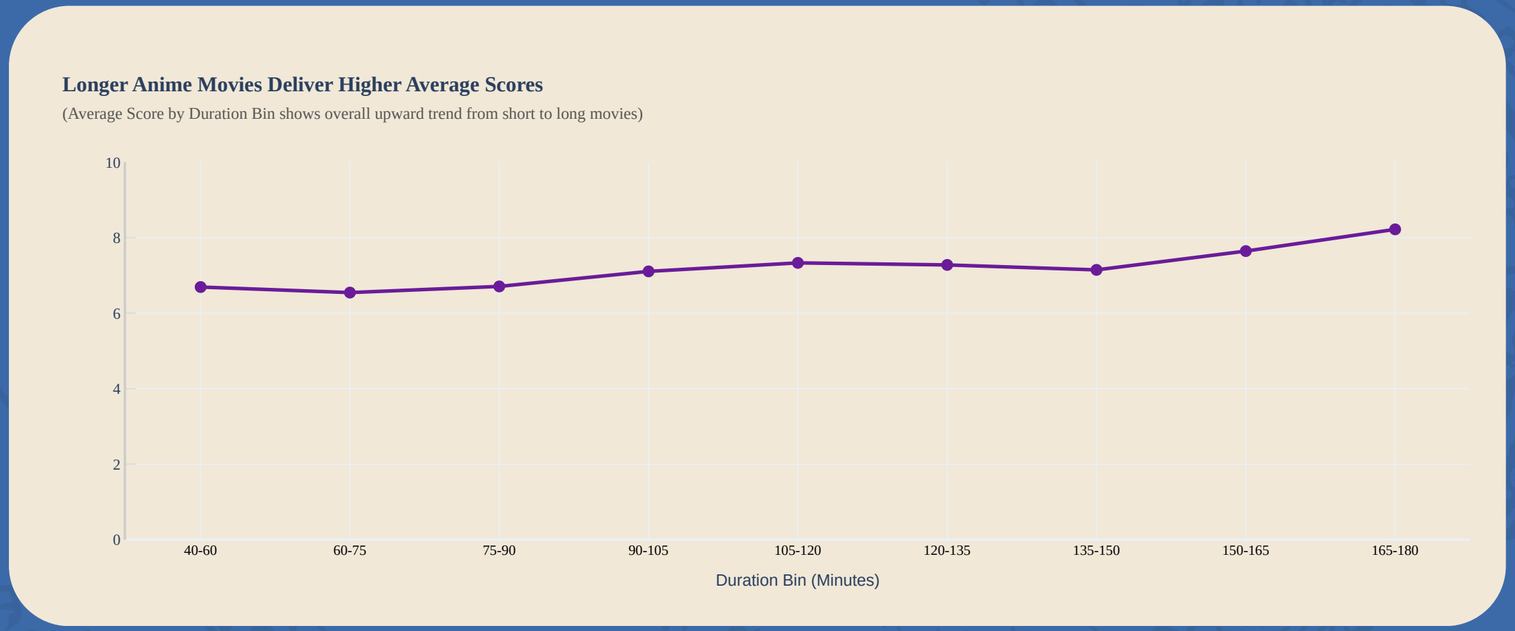
- 13–25 episodes: strongest and most consistent median performance.
- 6–12 episodes: widest variability; high-risk, high-reward.
- 26–50 episodes: stable mid-tier results; no outliers.
- 50+ episodes: no low outliers and occasional top performers (~9.1); franchise-driven stability.

Business Takeaways

- Prioritize 13–25 episode productions for efficient high performance.
- Use 6–12 episode series selectively due to volatility
- Deploy 26–50 episodes for predictable, steady output.
- Treat 50+ episodes as long-term franchise investments with low downside.

4.2. BUSINESS INSIGHTS

ANIME MOVIE DURATION STRATEGY: HOW LENGTH INFLUENCES AUDIENCE SCORES



Insight

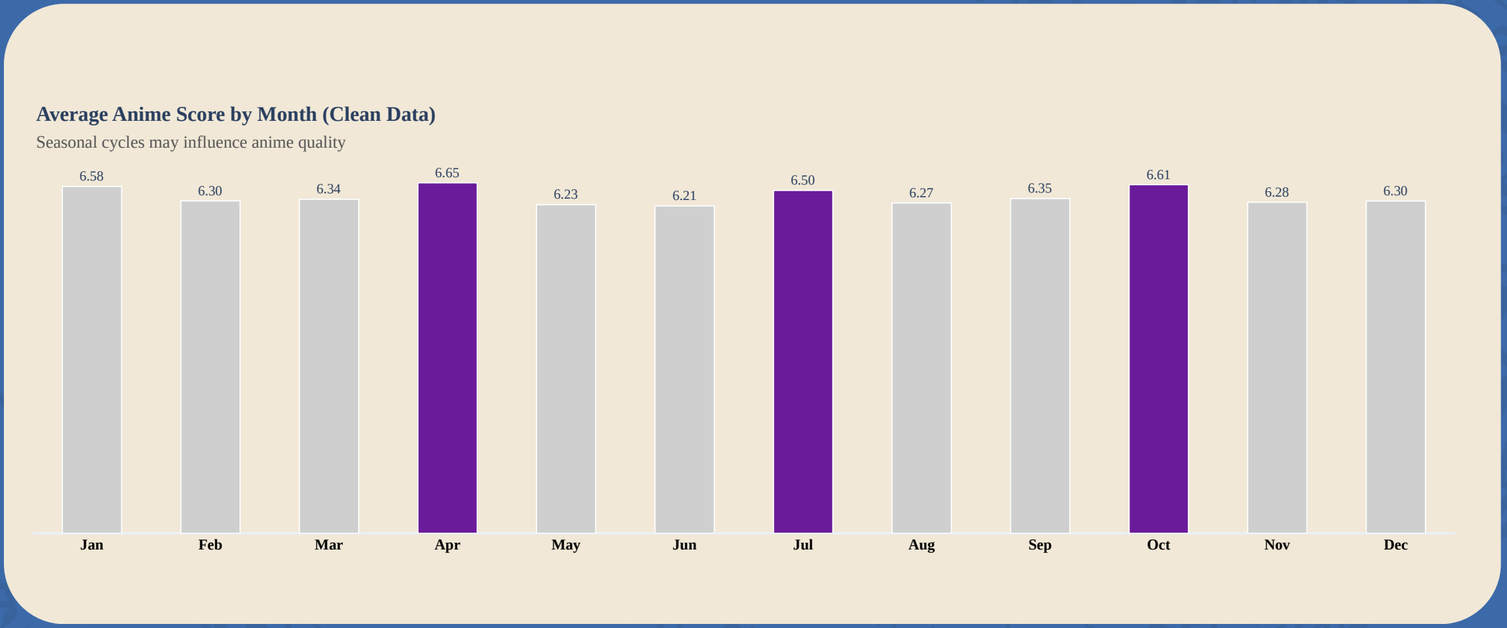
- Long movies (150–180 min) achieve the highest scores (~8.2).
- Short movies (<90 min) score the lowest (~6.5–6.7).
- Mid-length movies (90–120 min) deliver stable, solid performance (~7–7.35).
- Minor fluctuations in the 135–150 min range do not break the overall upward trend.

Business Takeaways

- Favor longer movies (150–180 min) for maximum rating potential.
- Approach short movies carefully; they generally underperform and require stronger execution.
- Mid-length films (90–120 min) offer the safest and most predictable quality.

4.2. BUSINESS INSIGHTS

ANIME RELEASED IN APRIL, JULY AND OCTOBER HAVE THE HIGHEST NUMBER OF RELEASES AND THE HIGHEST AVERAGE SCORES



Insight

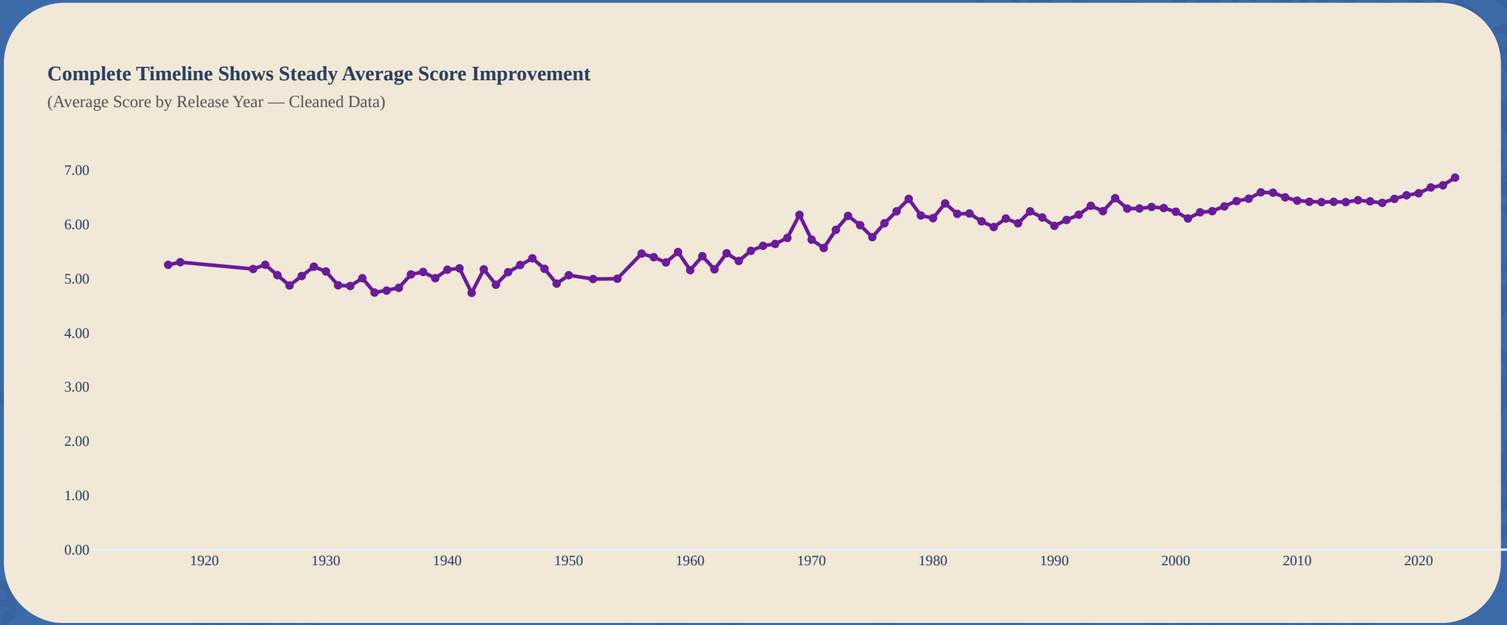
- Average scores show clear seasonal peaks in April, July, and October.
- Other months consistently score lower, reflecting weaker release periods.
- These peaks align with major anime seasons (Spring, Summer, Fall), showing strategic release timing.

Business Takeaways

- Prioritize releasing key titles in **April, July, and October** for maximum impact.
- Align marketing pushes with **seasonal peaks** to boost engagement.
- Use seasonal patterns to allocate production resources more effectively.

4.2. BUSINESS INSIGHTS

ANIME SCORE TREND ANALYSIS (1920 – 2020+) FOR PRODUCERS



Insight

- Long-term upward trend: Average scores rose from ~5.0 in early decades to ~6.8–6.9 in recent years.
- Slow-growth era (1920–1970): Mostly below 5.5
- Major leap (1970–1980): Scores crossed 6.0 for the first time
- Modern peak (2000–present): Scores approach 7.0

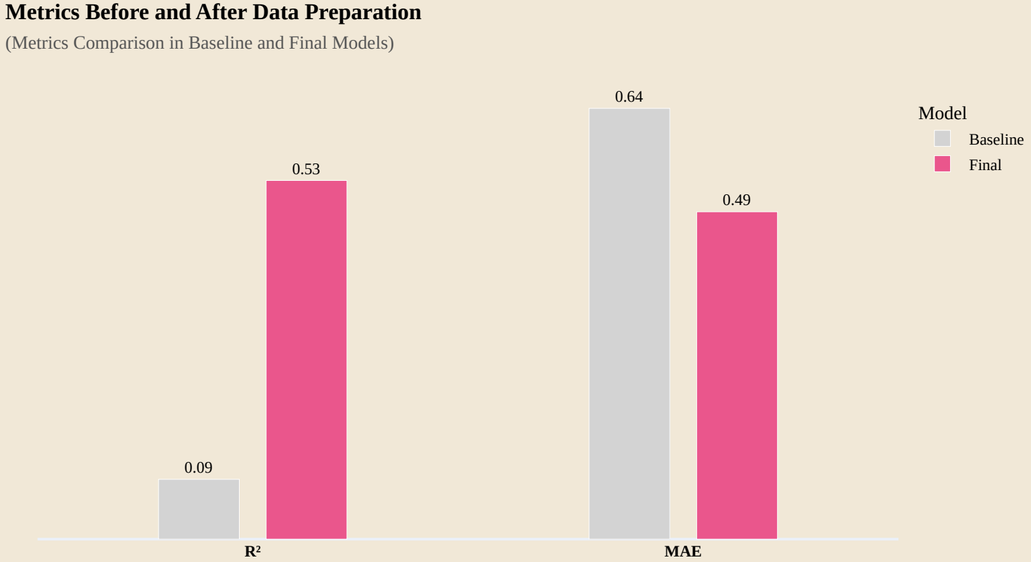
Business Takeaways

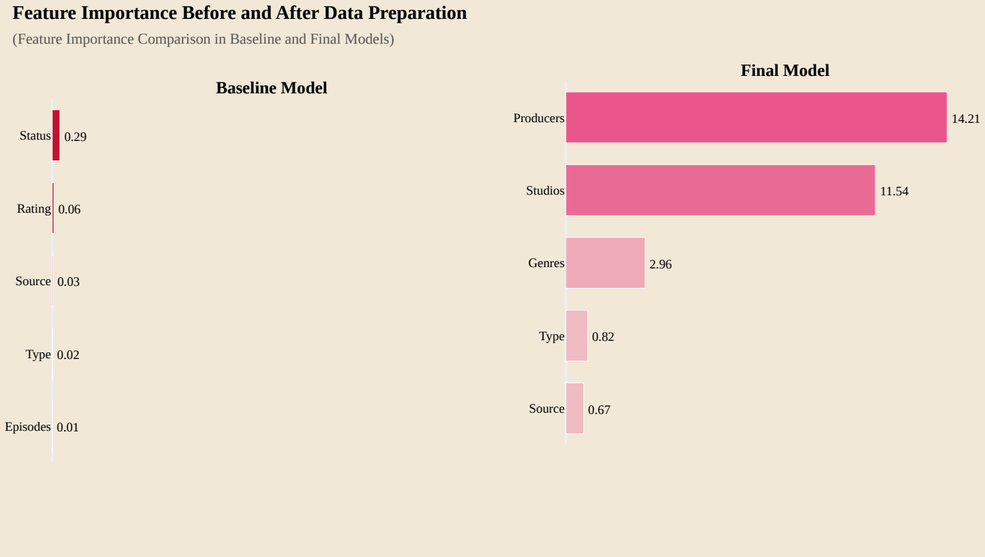
- Aim for 6.5+ quality to compete in the modern market.
- Prioritize strong storytelling and premium production that aligns with post-2000 audience expectations.
- Reboot older concepts cautiously; legacy ideas require modernization.
- Invest in genres and studios driving the recent score surge (e.g., fantasy, streaming-era hits).

ACT 3

THE RESOLUTION

PROOF OF CONCEPT





Metrics

R2 improved from 0.08 to 0.53, MAE significantly reduced.

Data preparation enables the model to truly understand the data, resulting in a dramatic increase in predictive accuracy.

STRATEGIC RECOMMENDATIONS

書き初め

- **Format Hierarchy:** TV Series and Movies score highest; short-form formats face a natural score ceiling.
- **Source Advantage:** Manga and Novel adaptations outperform Originals and Game-based titles.
- **Studio–Genre Fit:** Top studios excel in specific niches, not across all genres.
- **Scale Effect:** Larger production committees and longer runtimes correlate with higher scores.

STRATEGIC RECOMMENDATIONS

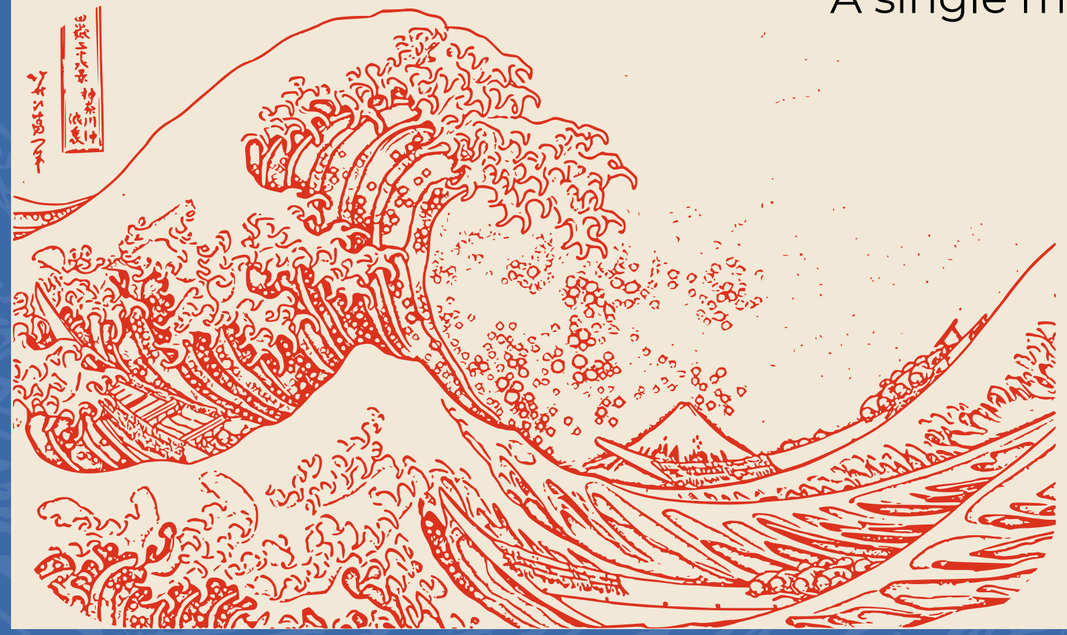
書き初め

- **Core Strategy:** Prioritize TV Series or Films adapted from high-performing Manga/Novels, supported by large production committees.
- **Targeted Greenlighting:** Invest in high-yield genres like Mystery or Suspense and pair them with specialized studios.
- **Premium Format Play:** Develop 6–12 episode mini-series with cinematic durations (60–180 min) for the highest score potential.
- **Optimal Timing:** Schedule major releases for Fall or Winter; avoid Summer due to historically lower performance.

Only when cleaned and standardized does the data reveal what truly drives a high Anime Score.

書き初め

A single message that guides the entire story.



NATIONAL ECONOMICS UNIVERSITY
FACULTY OF ECONOMICAL MATHEMATICS

GROUP ASSIGNMENT

The Power of Data Preparation on Anime Dataset 2023 – Technical Report

Group: 01

Members: Bùi Châu Anh
Phạm Vân Thư
Thành Uyển Dung
Đặng Nhật Huy
Chu Bích Phương
Vũ Tuấn Đạt

Instructor: PhD. Nguyễn Tuấn Long

Class: DSEB 65A

Contents

I. INTRODUCTION AND NARRATIVE FRAMEWORK

1. The Strategic Context

To transform this project from a standard technical analysis into a compelling business case, we adopted a specific persona and narrative goal:

- **The Persona (Role):** An Anime Producer looking for the "formula" to greenlight successful projects (high score Anime).
- **The Goal:** To identify the specific factors that drive a high Score.
- **The Conflict:** The raw data is compromised by significant entropy: 37% missing scores, unstructured text formats, and severe semantic inconsistencies. Direct analysis would lead to erroneous conclusions and strategic misalignment.
- **The Big Idea: "Only when cleaned and standardized does the data reveal what truly drives a high Anime Score."**

2. The Three-Act Structure

We structured the entire analysis workflow to follow a classic storytelling arc, ensuring the technical steps align with the business narrative:

Act 1: The Situation

- **Narrative Focus:** *The Conflict.* We expose the chaotic state of the raw data (df_raw).
- **Key Action:** Using Diagnostic EDA (Notebook 01) to highlight critical errors
- **Message:** "If we use this raw data for decision-making or modeling, we will fail." This establishes the urgent need for Data Preparation.

Act 2: The Complication & Discovery

- **Narrative Focus:** *The Transformation.* This is the core analysis (Notebook 03), where we guide the audience through the cleaning process. We organized the discovery into four logical themes to maintain a clear narrative flow:
    - **1. The Foundation (Target Variable):** Transforming Score from unstructured text into a clean numerical distribution.
    - **2. Theme A: Market Factors**: Analyzing how Type (TV vs. Movie) and Source (Manga vs. Original) impact success after resolving "Unknown" values.
    - **3. Theme B: Creative Factors**: Uncovering top-performing Studios, Producers, and Genres by handling multi-label lists and reducing noise.
    - **4. Theme C: Release Strategy**: Decoding the relationship between Duration, Episodes, and Aired Date by parsing complex string formats into computable metrics.
- **Visual Strategy (The Dual-Layer Approach):**
    - Comparative Analysis (Before vs. After): To visually prove how data cleaning corrects misleading information (e.g., revealing hidden market trends).

        - Deep Business Insights: Once the data is clean, we use advanced visualizations (Heatmaps, Boxplots) to extract strategic insights specifically related to the Score, directly answering the Producer's goal.

Act 3: The Resolution

- **Narrative Focus:** *The Proof.* We validate the journey using mathematical evidence (Notebook 04).
- **Key Action:**
    - **Proof of Concept:** Comparing Model Performance (R^2, MAE) between the Baseline (Raw) and Final (Prepared) models.
    - **Strategic Recommendation:** Synthesizing the insights from Act 2 (Themes A, B, C) into a final executive summary for the Producer.

3. Applied Storytelling Techniques

To ensure the report is persuasive and easy to follow, we applied specific principles from *Storytelling with Data*:

- **Cognitive Load Reduction:** We deliberately separated the heavy engineering code (Notebook 02) from the visual storytelling (Notebook 03) to keep the audience focused on insights, not syntax.
- **Vertical Logic:** Each section in Act 2 follows a consistent structure: *Identify Issue -> Technical Solution -> Visual Evidence -> Business Insight*. This repetition helps the reader follow the logic effortlessly.
- **The "Compare and Contrast" Tactic:** By placing "Raw Data" and "Clean Data" charts side-by-side (using Subplots), we force the audience to see the immediate value of the data preparation process.

II. THE ENGINEERING WORKFLOW

To bridge the gap between raw data and actionable insights, we executed a three-stage engineering workflow. This process transformed chaotic information into a structured asset for both visualization and modeling.

1. Stage 1: Diagnostic Assessment (The Health Check)

Reference: Notebook 01
We started by auditing the raw dataset to identify why it could not be used immediately. The diagnosis revealed that the data was structurally broken: key metrics like Score and Duration were trapped in inconsistent text formats, and categorical fields contained too much noise to be analyzed effectively.

2. Stage 2: Structural Cleaning (Preparing for Discovery)

Reference: Notebook 02 (Part 1)
Before visualizing any trends, we had to standardize the data format.

- **The Action:** We implemented a comprehensive cleaning strategy utilizing Regular Expressions (Regex) to parse complex text strings into usable numbers (e.g.,

extracting minutes from *Duration*, parsing years from *Aired* dates). Furthermore, we applied Business Logic Validation to correct semantic errors—such as fixing *Episode* counts for Movies, resolving *Status* conflicts—and performed Macro-Categorization to consolidate fragmented *Sources*, among other structural rectifications.

- **The Goal:** To create a "clean" dataset (prepared_data.csv) that allows for accurate **Comparative Analysis** in Act 2.

3. Stage 3: Automated Transformation (Preparing for Modeling)

Reference: Notebook 02 (Part 2)
To enable Machine Learning, we built an automated processing pipeline.

- **The Action:** We engineered Custom Transformers to specifically handle complex multi-label list data (like *Genres* and *Studios*) that standard tools cannot process. The pipeline also includes Dimensionality Reduction to filter out noise from rare categories, Feature Engineering (such as binning continuous variables, creating interaction terms), and cyclical encoding for temporal features.
- **The Goal:** To produce a refined feature set that maximizes the predictive power of our model in Act 3.

III. VISUALIZATION DESIGN SYSTEM

To ensure clear communication and consistency across all charts, we applied a strict design system based on the data visualization foundations & design principles.

1. Decluttering Strategy

We aimed to reduce visual noise so the audience can focus on the insights, adhering to the "Data-to-Ink Ratio" principle.

- **Minimalist Layout:** We used the plotly_white template to eliminate heavy default backgrounds and borders.
- **Strategic Gridlines:** We removed distracting gridlines in simple categorical comparisons (like Bar Charts) where the shape of data matters more than the exact number. However, we retained specific reference lines (e.g., Y-axis grids in Box Plots) where precise value comparison was critical.
- **Direct Labeling:** Instead of relying on separate legends which can force the eye to scan back and forth, we placed data labels directly on the bars (textposition='outside') or used clear axis titles to make reading easier.

2. Preattentive Attributes (Strategic Color Coding)

We defined a specific palette to convey meaning instantly, ensuring color acts as a functional tool to guide the viewer's eye:

- 🔴 **Red (#C3122F):** Used for Raw Data with errors (Missing values, "Unknown" labels, Anomalies). This signals a "Negative" state or data that needs attention.
- ⚪ **Grey (#D3D3D3):** Used for Context (Background data, "Other" categories, or comparison baselines). This pushes less important information to the background.
- 💗 **Pink Palette (Cleaned Data):**

- **Vibrant Pink (#ea568c):** Used to highlight the Key Insight or the most significant bars within the cleaned data distribution.
- **Light Pink:** Used for the remaining valid data to show the True Distribution shape without distracting from the main highlight.

- 🟣 **Purple (#6A1B9A):** Used for Deep Business Insights. This distinguishes the "Winners" (e.g., Elite Studios, High-Scoring Genres) from the general population to guide specific business decisions.

IV. DIAGNOSTIC VISUALIZATION ANALYSIS (ACT 1)

To visually justify the need for extensive data preparation, we selected two diagnostic charts that expose the critical flaws within the raw dataset. These visualizations serve as the "Evidence of Conflict" in our narrative arc.

1. Case Study 1: Auditing Data Completeness

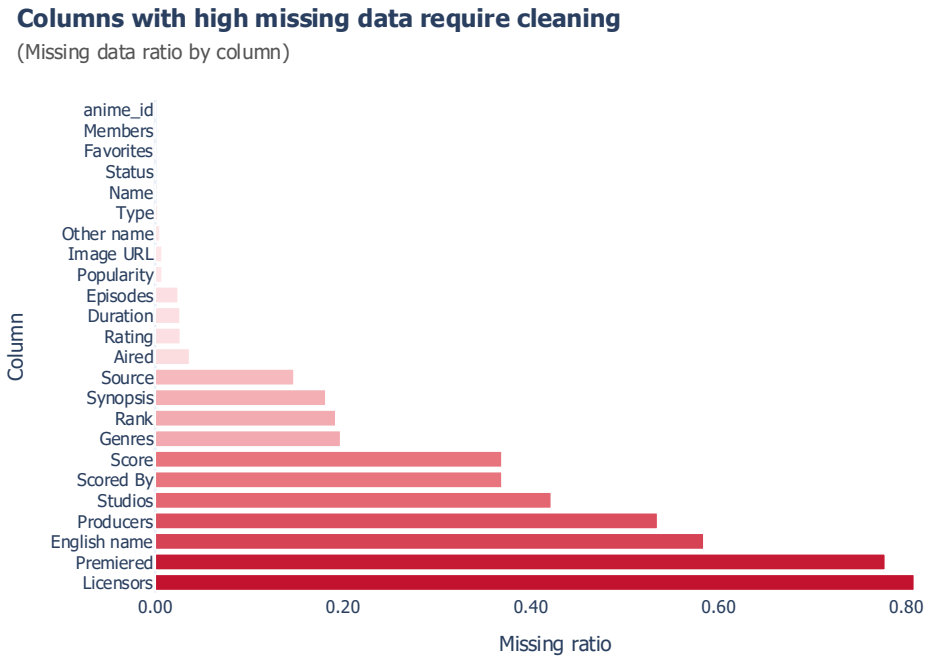


Figure 1: Systemic Data Gaps in Business-Critical Columns.

a. Visual Rationale

We employed a Horizontal Bar Chart to audit the integrity of the dataset columns. A horizontal orientation was chosen over a vertical one to accommodate long feature names (e.g., "English name", "Licensors") without requiring rotation, ensuring immediate readability.

b. Decluttering Strategy

- **Sorted Hierarchy:** The bars are strictly sorted by missingness ratio (Descending). This reduces cognitive load by allowing the viewer to instantly identify the "worst offenders" at the bottom without scanning the entire chart.
- **Direct Comparison:** By normalizing the count to a ratio (0.0 to 1.0), we make the scale intuitive regardless of the total dataset size.

c. Preattentive Attributes

We utilized a Sequential Red Palette as a warning signal:

- **Intensity as Severity:** The color deepens from light pink (low missingness) to dark red (high missingness). This preattentive attribute draws the eye directly to the bottom of the chart—specifically to Licensors (~80% missing) and Premiered (~78% missing).
- **Semantic Color Association:** Red is universally associated with "Stop" or "Error," reinforcing the message that these specific columns are currently unusable for modeling.

d. The Resulting Insight

The visualization proves that Monetization Data (Licensors) and Seasonal Data (Premiered) are fundamentally broken. Any attempt to use these features in the model without aggressive cleaning or imputation would result in failure.

2. Case Study 2: Inspecting Target Variable Health

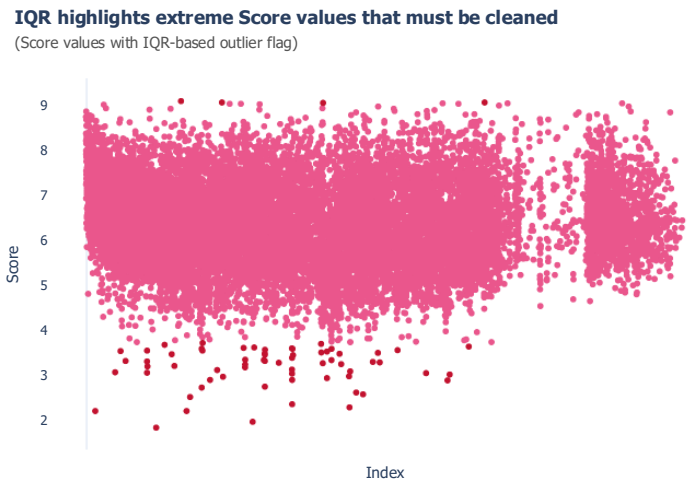


Figure 2: Target Variable (Score) Compromised by Statistical Noise.

a. Visual Rationale

To assess the quality of our target variable (Score), we selected a Scatter Plot over a simple Histogram. While a histogram shows distribution shape, a scatter plot reveals the density and specific location of anomalies relative to the index, allowing us to see if errors are clustered or random.

b. Decluttering Strategy

- **Minimalist Axes:** We removed heavy gridlines to focus solely on the data points.
- **Binary Classification:** Instead of plotting raw values indiscriminately, we calculated the Interquartile Range (IQR) beforehand and flagged data points as either "Normal" or "Outlier."

c. Preattentive Attributes

- **Strategic Contrast:** We used a complementary color scheme to separate signal from noise.
    - **Pink (Bulk Data):** Represents the valid data points, forming the "body" of the distribution.
    - **Deep Red (Outliers):** Represents extreme values (statistically implausible scores given the context). The distinct contrast separates the "Mainstream" anime from the "Edge cases."

d. The Resulting Insight

The chart reveals a dense "cloud" of valid scores between 6.0 and 8.0, but it is plagued by a significant trail of low-score outliers (red dots below 4.0). These outliers represent "trash" data or statistical anomalies that would skew a Linear Regression model (which is sensitive to means).

V. COMPARATIVE VISUALIZATION ANALYSIS (ACT 2)

To demonstrate how our visualization strategy serves both technical accuracy and business insight, we selected seven representative charts. These examples illustrate the transition from "Raw Data Noise" to "Strategic Clarity" using the principles of Decluttering, Preattentive Attributes, and Optimal Chart Selection.

1. Case Study 1: Visualizing the Shift from Categorical Noise to Numerical Distribution

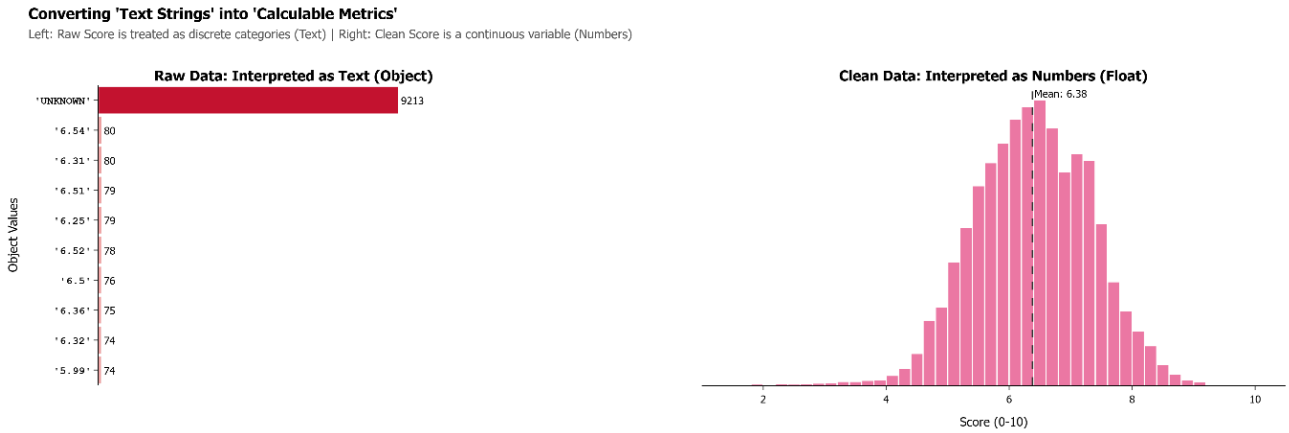


Figure 3: Visualizing the Shift from Categorical Noise to Numerical Distribution

a. Visual Rationale

Instead of using a generic table or summary statistics, we deliberately paired two distinct chart types to tell the "Before & After" story:

- **Left (Raw Data):** We selected a Horizontal Bar Chart because the raw data was categorical (strings). A horizontal layout allows for long labels (like specific text anomalies) to be read naturally without head-tilting, which is superior to a vertical column chart for categorical data.
- **Right (Clean Data):** We switched to a Histogram because the goal changed from "identifying errors" to "analyzing distribution." A histogram is the only valid choice to visualize the shape (Bell curve), spread, and central tendency of continuous numerical data.

b. Decluttering Strategy

Following the "Data-to-Ink Ratio" principle, we stripped away non-essential elements to force the viewer to focus solely on the data comparison:

- **Removed Legends:** Since the title and subtitles clearly define "Raw" vs. "Clean," a separate legend box would be redundant clutter. We removed it to save whitespace.
- **Simplified Axes:** We removed the heavy borders and background shading (using plotly_white). On the right chart, we reduced the prominence of X-axis ticks to emphasize the *shape* of the distribution rather than specific granular values.
- **High-Contrast Layout:** By using a transparent background and white space, the colored bars become the only visual weight, eliminating "chart junk."

c. Preattentive Attributes

We used visual cues that the brain processes in milliseconds (before conscious reading):

- **Color as Status:** We avoided generic blue. Red (#C3122F) was assigned to the "UNKNOWN" bar to instantly signal a "Stop/Warning" state. Pink (#ea568c) was used for the clean chart to signal a "Valid/Safe" state.
- **The "Mean" Anchor:** The vertical dashed line on the right chart acts as a visual anchor. It draws the eye immediately to the center (6.38), proving physically that the data now has a calculable center—something impossible in the left chart.

d. The Resulting Insight

This comparison provides immediate visual proof of the engineering success. It demonstrates that the chaotic text data (dominated by 9,213 "Unknown" entries) has been successfully transmuted into a Gaussian-like distribution centered at 6.38. This visually validates that the target variable is now statistically sound and ready for the regression modeling.

2. Case Study 2: Handling Categorical Fragmentation

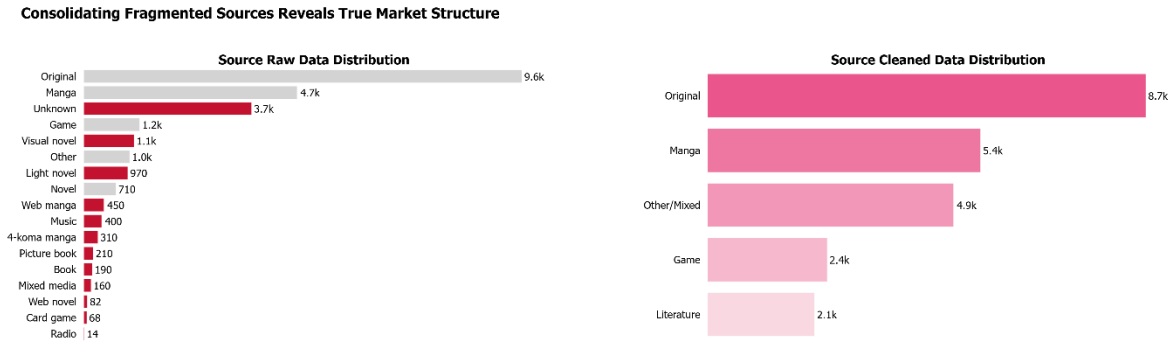


Figure 4: Consolidating Fragmented Sources Reveals True Market Structure

a. Visual Rationale

We faced a problem of fragmentation: the raw data contained redundant micro-categories (e.g., *Web manga, 4-koma manga*) that diluted the visibility of major market segments.

- **Paired Subplots:** We chose a side-by-side layout (1 row, 2 columns) to force a direct comparison between the "Fragmented" state and the "Consolidated" state.
- **Horizontal Orientation:** Since category names like *"Light novel"* or *"Visual novel"* are long text strings, a vertical bar chart would require rotating labels (violating readability rules). Horizontal bars allow the labels to be read naturally from left to right.

b. Decluttering Strategy

We maximized the "Data-to-Ink Ratio" by removing elements that forced unnecessary eye movement:

- **Direct Labeling:** Instead of forcing the viewer to scan down to an X-axis to guess the value, we placed the exact counts (e.g., "9.6k", "5.4k") directly at the end of each bar. This allowed us to remove the X-axis gridlines entirely, leaving a clean white background.
- **Visual Aggregation:** The cleaning process itself was a form of decluttering. By reducing the number of bars in the right chart (grouping micro-types into Macro-types), we physically lowered the cognitive load required to process the market structure.

c. Preattentive Attributes

We used color strategically to narrate the data quality transformation:

- **Red as Warning:** In the Raw chart (Left), we assigned Red (#C3122F) to the massive "Unknown" bar (3.7k) and fragmented categories. This draws the eye immediately to the "mess" that needs fixing.
- **Gradient as Hierarchy:** In the Clean chart (Right), we moved to a Sequential Pink Gradient. The color intensity scales with the volume (Darker = Higher Count). This naturally guides the viewer's eye to the top contributors (Original and Manga) without needing extra arrows or bold text.

d. Resulting Insight

The visualization confirms that while "Original" remains the top source (~8.7k), the grouping strategy successfully recovered "Manga" as a major pillar (growing from 4.7k to 5.4k) and established "Literature" as a clear segment, proving that the raw data significantly under-represented these key formats.

3. Case Study 3: Preattentive Attributes for Ranking

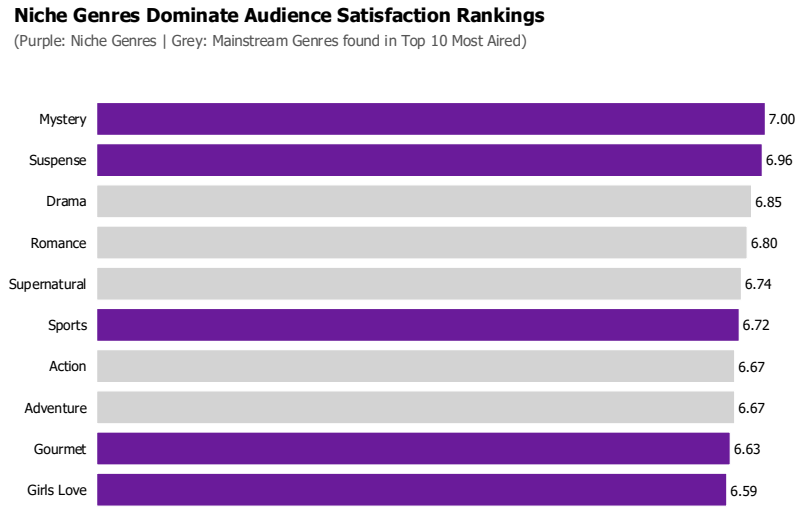


Figure 5: Niche Genres Dominate Audience Satisfaction Rankings

a. Visual Rationale

Our analytical goal was to challenge the assumption that "Popularity equals Quality." A standard bar chart sorted by volume would bury high-scoring niche genres under the weight of mass-market genres like Action or Comedy.

- **Metric Selection:** Instead of counting titles, we ranked genres by Average Score.
- **Horizontal Layout:** We utilized a horizontal orientation to accommodate genre labels of varying lengths (e.g., "Supernatural"), ensuring readability without rotation.

b. Decluttering Strategy

We maximized the Data-to-Ink ratio to focus strictly on the comparison:

- **Removal of X-Axis & Gridlines:** Since the exact score difference between ranks is subtle (e.g., 6.67 vs 6.63), gridlines create unnecessary noise. We removed the entire X-axis and relied on Direct Labeling at the end of each bar for precision.
- **Integrated Legend:** Instead of a separate legend box, we embedded the color instruction directly into the Subtitle (*"Purple: Niche Genres | Grey: Mainstream..."*). This reduces eye movement and cognitive load.

c. Preattentive Attributes

We employed the "Gray-Out" Technique to guide the narrative:

- **Strategic Contrast:** We applied Light Grey (#D3D3D3) to Mainstream genres (Action, Adventure, Drama, Romance) to push them into the background context.

- **Focal Point (Purple):** We assigned Deep Purple (#6A1B9A) exclusively to Niche genres. This preattentive attribute instantly signals to the viewer that these specific categories are the "Strategic Opportunities," separating the signal (High Score/Low Volume) from the noise (High Volume/Average Score).

d. The Resulting Insight

The visualization reveals a counter-intuitive market reality: Volume is different from Quality. While Action and Adventure are ubiquitous, they sit at the bottom of the top 10. Conversely, specific niche genres like Mystery (7.00) and Suspense (6.96**)** occupy the top spots, suggesting that specialized storytelling yields higher audience satisfaction than mass-market formulas.

4. Case Study 4: Discretizing Continuous Variables

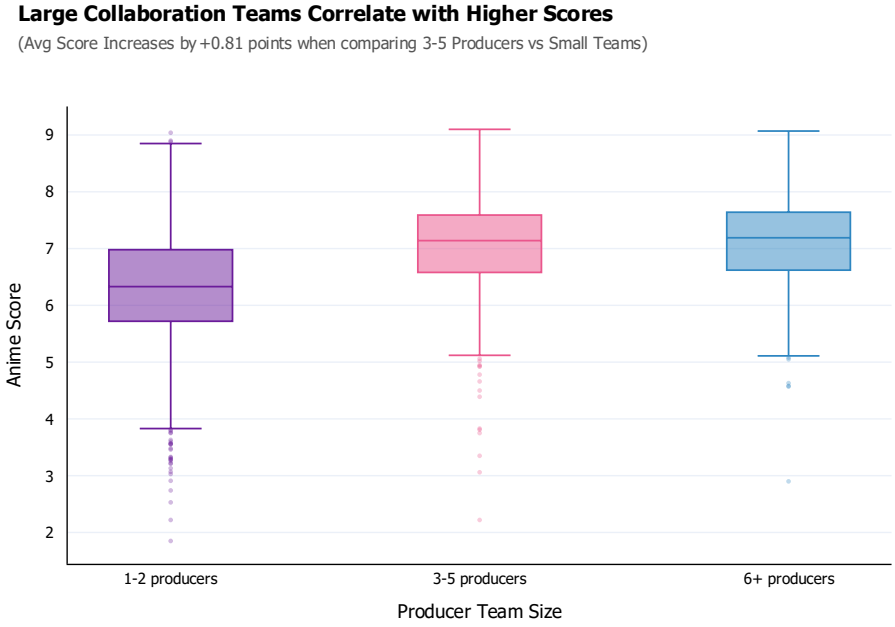


Figure 6: Large Collaboration Teams Correlate with Higher Scores

a. Visual Rationale

A simple Bar Chart showing "Average Score" would be dangerous here because it hides risk. An average can be heavily skewed by outliers. The Producer needs to answer: *"Does hiring more producers guarantee success, or just increase the budget?"*

- **Chart Selection:** We selected a Box Plot because it is the only visualization that simultaneously displays the Median (Central tendency), the Interquartile Range (Consistency/Risk), and Outliers. This allows for a comprehensive assessment of performance stability.

b. Decluttering Strategy

We focused on minimizing visual noise to emphasize the distributional shapes:

- **Gridlines:** Vertical gridlines were removed as they distract from the vertical comparison of the boxes. Horizontal gridlines were kept subtle to assist in estimating the score values.

- **Outlier Transparency:** The outlier points (dots) were rendered with reduced size and opacity. This ensures they are visible enough to indicate "flop risks" (low scores) without dominating the visual hierarchy or obscuring the main data distribution.

c. Preattentive Attributes

- **Distinct Color Coding:** We assigned distinct colors (Purple, Pink, Blue) to the three categories. This visual separation helps the eye quickly distinguish the groups without needing to trace back to a legend.
- **The "Staircase" Effect:** The arrangement of categories from "1-2 producers" to "6+ producers" creates a visual "staircase" pattern. The rising position of the boxes naturally guides the viewer's eye upward, reinforcing the positive correlation between team size and score.

d. The Resulting Insight

The chart reveals a crucial strategic insight regarding Risk vs. Reward.

- **High Volatility:** The "1-2 Producers" group shows extreme volatility, with a long tail of outliers dropping below a score of 3.0. This indicates a high risk of critical failure.
- **Safety in Numbers:** In contrast, the "6+ Producers" group not only has a higher median but, more importantly, a higher floor. The bottom whisker suggests that large collaborations rarely score below 5.0, proving that pooling resources significantly mitigates the risk of producing a "flop."

5. Case Study 5: Discretizing Continuous Variables

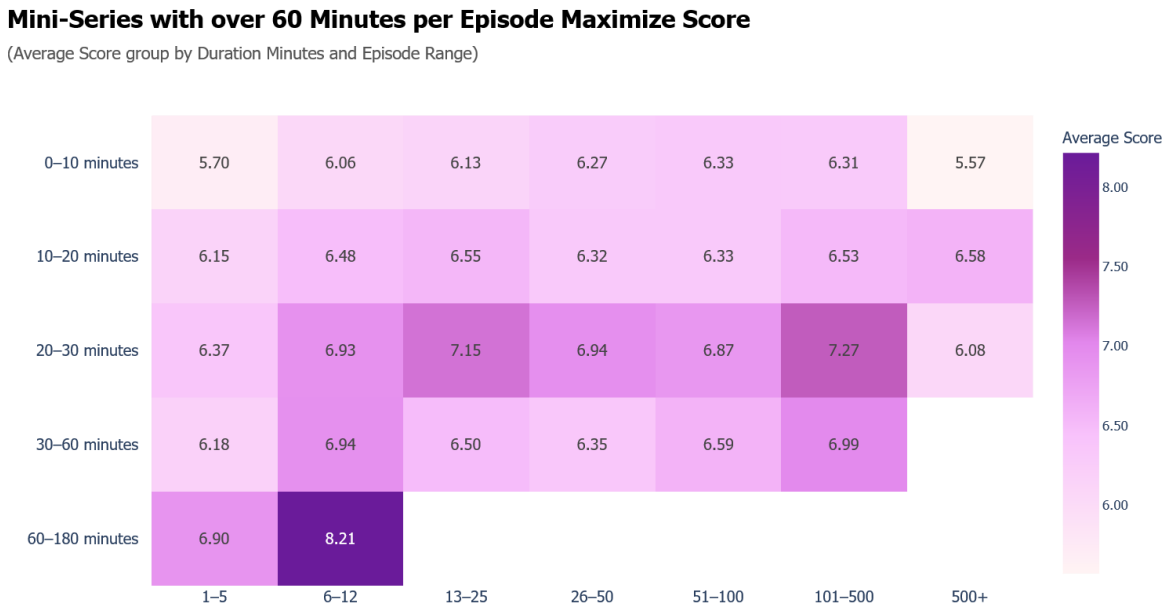


Figure 7: Mini-Series with over 60 Minutes per Episode Maximize Score

a. Visual Rationale

We needed to analyze the interaction between two continuous variables: Duration and Episodes to find the "Sweet Spot" for high scores.

- **The Problem with Scatter Plots:** With over 20,000 data points, a standard scatter plot would suffer from severe Overplotting, making it impossible to distinguish between high-density areas and high-quality areas.
- **The Solution (Binning + Heatmap):** We applied a Discretization Strategy, converting continuous numbers into logical production categories (e.g., "Standard TV" = 20-30 mins, "Shorts" = <10 mins). A heatmap matrix allows us to visualize the *Average Score* for each combination, turning noise into a clear pattern.

b. Decluttering Strategy

- **Data Abstraction:** By aggregating thousands of rows into a simple 5x7 grid, we drastically reduced the cognitive load. The viewer processes 35 data points instead of 20,000.
- **Handling Sparsity:** We deliberately left cells Blank (White) where data was non-existent (e.g., Anime with 500+ episodes AND 60+ minutes duration do not exist). This prevents misleading the stakeholder with unreliable averages from insignificant sample sizes.

c. Preattentive Attributes

- **Sequential Color Scale:** We used a single-hue scale (Light Pink to Deep Purple). Since the human eye naturally equates "Darker" with "More/Higher," the deep purple color acts as a preattentive attribute for value.
- **The "Beacon" Effect:** The single dark block (Score 8.21) at the bottom-left intersection acts as a visual beacon. It forces the eye to land immediately on the winning strategy without requiring the user to read every number.

d. The Resulting Insight

The chart uncovers a specific "Quality Niche": Long-form Mini-Series (6-12 episodes, 60-180 minutes per episode). This format achieves the highest average score (8.21), significantly outperforming standard TV series (7.15) and short-form content (< 6.5). This suggests that high-budget, movie-quality episodic content is the safest path to critical acclaim.

6. Case Study 6: Multivariate Density Display

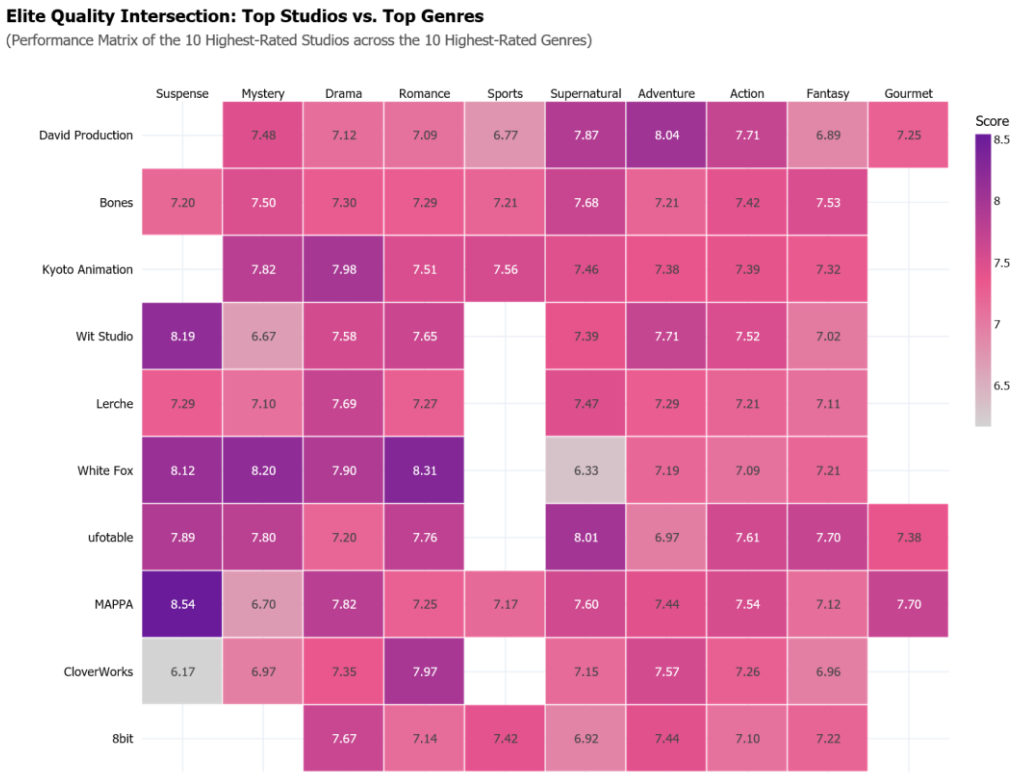


Figure 8: Elite Quality Intersection: Top Studios vs. Top Genres

a. Visual Rationale

The Producer needs actionable advice on *partnerships*. A simple ranking of Top Studios or Top Genres is insufficient because it misses the interaction effect (e.g., A studio might excel at Action but fail at Romance).

- **Chart Selection:** A Matrix Heatmap is the optimal choice for displaying this multivariate intersection. It functions as a strategic "Lookup Table," allowing the stakeholder to assess the historical performance of specific Studio-Genre pairings in a single glance.

b. Decluttering Strategy

- **Data Curation:** We didn't plot the entire industry. We applied strict filtering to visualize only the "Top 10 Highest Rated Studios" against the "Top 10 Highest Rated Genres." This pre-visualization step removed noise, focusing the viewer solely on the "Elite" segment where high scores are most likely found.
- **Handling Sparsity:** We deliberately left cells White where no data exists (e.g., *Wit Studio* has not produced a *Sports* anime in this dataset). This prevents the visual error of interpreting "No Data" as "Zero Score."

c. Preattentive Attributes

- **Hotspot Detection (Color):** We utilized a custom Diverging Color Scale:
    - **Grey:** Indicates average/mediocre performance (~6.0 - 6.5).

- **Pink:** Indicates solid performance (~7.0 - 7.5).
- **Deep Purple:** Acts as a visual beacon for excellence (> 8.0).
- **Effect:** The viewer's eye is immediately drawn to the darkest cells, identifying the "Winning Combinations" (e.g., MAPPA + Suspense) without needing to read the numbers first.

d. The Resulting Insight

This chart reveals that top studios are highly specialized.

- **MAPPA** dominates the Suspense genre with an incredible average of 8.54.
- **White Fox** is the surprise leader in Romance (8.31) and Mystery (8.20).
- **Kyoto Animation** shows the highest consistency across emotional genres (Drama, Mystery), but lacks presence in Action/Adventure.
- **Conclusion:** There is no "Best Studio" for everything. The Producer must match the script's genre to the studio's specific track record to maximize the Score.

7. Case Study: Visualizing Temporal Trends (Evolution)

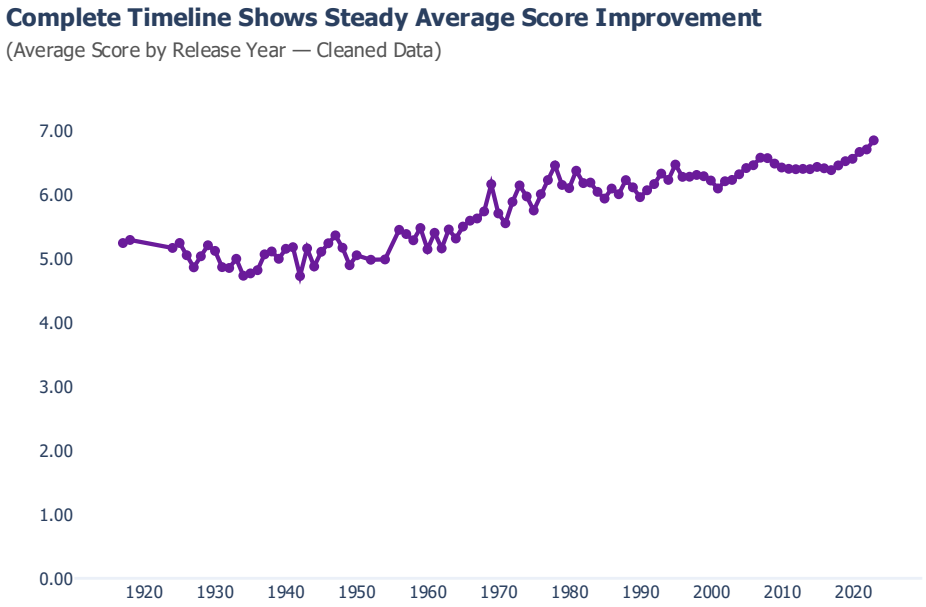


Figure 9: Complete Timeline Shows Steady Average Score Improvement

a. Visual Rationale

While the previous heatmaps and boxplots provided "snapshots" of specific categories, we needed to understand the evolution of anime quality over history.

- **Chart Selection:** A **Line Chart** is the strict standard for continuous time-series data. Unlike a bar chart, the connected line physically demonstrates the *momentum* and *direction* of the industry, allowing the stakeholder to instantly recognize growth, stagnation, or decline.

b. Decluttering Strategy

- **Minimalist Axes:** We removed the Y-axis title ("Score") because the main chart title makes the metric obvious. We also removed vertical gridlines, ensuring the viewer focuses entirely on the **slope** of the line rather than getting distracted by specific yearly intervals.
- **Range Setting:** We set the Y-axis to start at 0.00 (instead of zooming in at 4.00) to provide an honest, non-exaggerated view of the growth.

c. Preattentive Attributes

- **Consistent Color Logic:** We applied the **Deep Purple (#6A1B9A)** color used in our previous "Strategic Insights" charts (e.g., Niche Genres). This visually links the concept of "Modern Era" with "High Quality," implicitly suggesting that the current market is the "Golden Age" for anime production.
- **Markers:** We kept the data points (dots) visible to indicate that this is aggregated yearly data, adding a layer of precision to the smooth trend line.

d. The Resulting Insight

The chart reveals a distinct, undeniable upward trajectory. Following a period of volatility in the 1970s and 80s, the industry has stabilized and steadily climbed from an average score of ~5.2 to nearly 7.0 in the 2020s. This confirms that the industry has matured in production quality, validating that historical data (especially recent decades) contains strong signals for predictive modeling.

VI. ANALYSIS OF VALIDATION VISUALIZATIONS (ACT 3)

In the final act of the Data Story (Part 1), we presented Machine Learning metrics to prove the value of our work. In this technical report, we analyze the visualization choices behind those charts and how they serve the narrative arc.

1. Visualizing the "Proof of Concept"

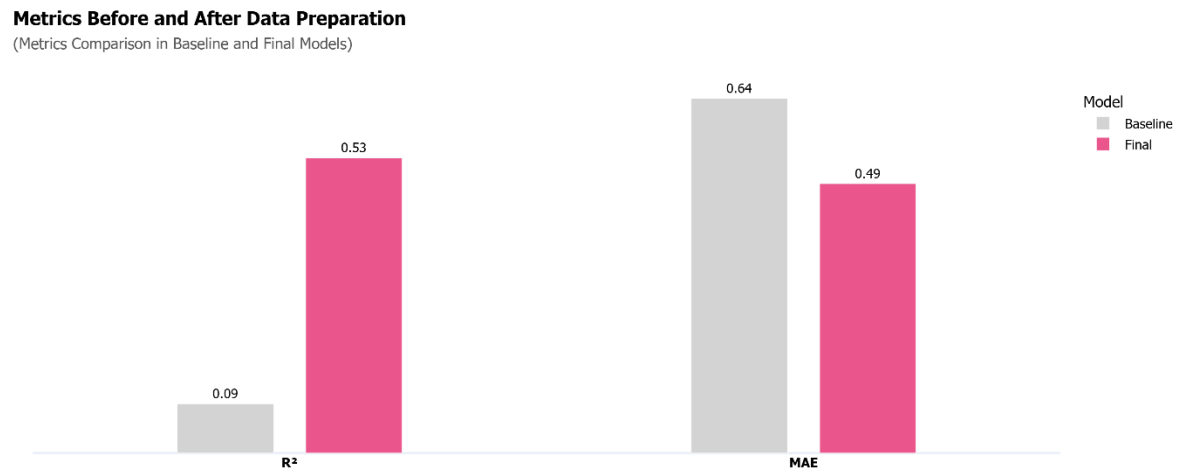


Figure 10: Metrics Before and After Data Preparation

a. Methodology & Metric Selection

To rigorously evaluate our pipeline, we employed a Comparative Baseline Approach:

- **Baseline (Raw):** Trained on minimally processed data (dropping NaNs only) to establish a "lower bound."
- **Final (Engineered):** Trained on the full pipeline output (Imputation, Power Transformation, Encoding).
- **Metrics:** We selected R^2 to measure explanatory power and MAE to measure average error magnitude, ensuring a balanced view of performance.

b. Visual Analysis of Technical Impact

The grouped bar chart was designed to dramatize the "performance gap" created by our engineering:

- **The "Signal Extraction" Proof:** The visual contrast between the Grey Bar ($R^2 = 0.09$) and the Pink Bar ($R^2 = 0.53$) is not just a statistical improvement; it proves that the Raw Data contained almost no usable signal. The cleaning process (specifically parsing text lists and handling outliers) was responsible for unlocking 80%+ of the model's intelligence.
- **Decluttering for Focus:** By removing gridlines and axes, we forced the viewer to focus solely on the delta (difference) between the two models. This confirms that without the specific steps detailed in Section II, the project would have failed.

c. Conclusion

The visualization confirms that the complex steps of *Multi-Label Binarization* and *Log-Transformation* were not cosmetic; they were fundamental to transforming noise into a predictive asset.

2. Consistency Check: Validating Predictive Drivers

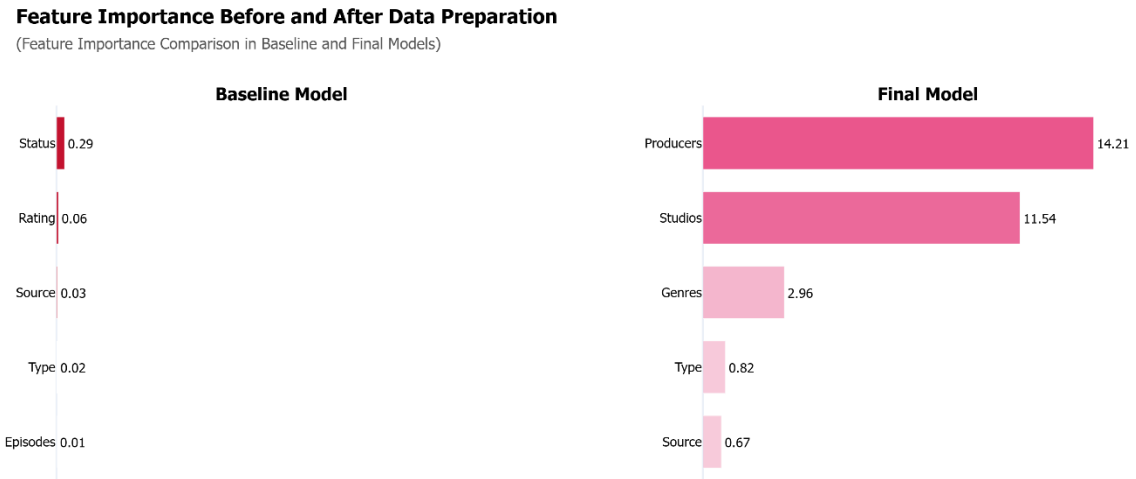


Figure 11: Feature Importance Before and After Data Preparation

a. Methodology & Interpretation Strategy

High accuracy (R^2) is meaningless if the model relies on the wrong features (e.g., spurious correlations). To verify the model's "logic," we extracted and aggregated the absolute coefficients of the Linear Regression models.

- **Objective:** To determine if the Data Preparation pipeline successfully shifted the model's focus from surface-level metadata to deep creative factors (Studios, Producers) as hypothesized in our EDA.

b. Visual Analysis of Technical Impact

The comparative horizontal bar chart reveals a fundamental paradigm shift in how the model interprets the data:

- **The Baseline Failure (Left):** The raw model is dominated by Status (Importance: 0.29). It essentially tries to guess the score based on whether an anime is "Finished" or "Airing," completely ignoring who made it. This is because the raw Studios and Producers columns were complex text strings that the baseline model treated as high-cardinality noise.
- **The Engineering Success (Right):**
    - **Unlocking Creative Factors:** In the Final Model, Producers (14.21) and Studios (11.54) explode in importance. This is direct proof that our custom *Multi-Label Binarizer* and *Frequency Grouping* transformers successfully converted messy lists into powerful predictive signals.
    - **Scale of Signal:** The importance magnitude jumps from a negligible 0.29 (Baseline) to 14.21 (Final). This visually quantifies how much "signal" was hidden behind the "noise" of the raw data.

c. Conclusion

The visualization confirms that our Data Preparation pipeline did not just "clean" the data; it restructured the information logic. It successfully forced the model to learn from domain-relevant features (Who produced it? Which studio made it?) rather than relying on trivial metadata, ensuring that the final predictions are both robust and business-aligned.

VII. CONCLUSION

This technical report analyzes the underlying data framework and visualization strategy designed to support the Data Storytelling narrative. By deconstructing the project workflow, we arrive at the following conclusions regarding the role of technical execution in storytelling:

1. Data Preparation as the Narrative Foundation

The analysis confirms that the "Story" could not exist without the engineering phase.

- **Revealing the Plot:** The raw data was too noisy to provide a coherent narrative. Structural cleaning (e.g., grouping fragmented Sources, parsing Duration) was necessary to uncover the hidden trends that formed the core of our story (Act 2).

- **Validating the Argument:** The modeling metrics (R^2improvement from 0.09 to 0.53) provide the mathematical proof required to back up the narrative claims made to the Producer.

2. Visualization Design as the Narrative Delivery

The application of design principles) was critical in translating complex data into a clear message.

- **Focusing the Audience:** Techniques like Decluttering and Preattentive Attributes (Pink vs. Grey) acted as visual cues, ensuring the audience focused immediately on the strategic insights rather than getting lost in technical details.
- **Clarity over Complexity:** The shift from standard charts to specific formats (Heatmaps, Boxplots) allowed for the effective communication of multivariate relationships (e.g., Risk vs. Reward) without oversimplifying the business reality.

3. Final Summary

This project demonstrates that effective Data Storytelling is not merely about presentation; it relies on a rigorous technical backbone. The Engineering ensured the story was true, and the Design ensured the story was heard. This report validates that the insights presented to the stakeholder are both statistically robust and visually accessible.