Data Cleaning & Insights Report for Video On-Demand Dataset

Prepared for Platform X

Oznur Ozcelikoglu | Data Analyst | MSc Data Science | BSc Mathematical Engineering

ozcelikogluoznur@yahoo.com | +44 7765 742516

Summary

This project involved cleaning and analyzing a three-month video-on-demand (VOD) dataset using Python. The main goal was to extract structured metadata—such as programme title, season, episode, and quality—from unstructured content names. A sequence of regex patterns was applied to parse the data, followed by fuzzy matching against a master lookup table for validation. The cleaned dataset was then used to generate insights, including top-performing programmes, monthly viewership trends, and episode-level audience patterns. This end-to-end pipeline demonstrates a robust and scalable approach to handling real-world content metadata and deriving meaningful insights from it.

Contents

Figures Table

Data Cleaning & Insights Report for Video On-Demand Dataset

1. Introduction

This report presents the results of a data cleaning and analysis task conducted on a three-month sample of video-on-demand (VOD) usage data. The task was assigned as part of a recruitment assessment and focuses on demonstrating the ability to transform raw, unstructured content data into a reliable and analyzable format using Python.

The dataset includes viewing records from Platform X, where programme details are embedded in a semi-structured “Content Name” field. These entries vary widely in format and may contain combinations of programme title, season, episode number, and video quality, requiring a careful parsing and cleaning strategy. A clean master lookup table was provided to validate and match parsed content accurately.

The objective of this report is twofold: (1) to outline the data cleaning methodology used to standardize and enrich the dataset, and (2) to highlight key insights derived from the cleaned data, such as top-performing programmes and viewing trends. The work emphasizes precision, automation, and interpretability in handling real-world media data.

2. Data Cleaning Approach & Methodology

Preprocessing

The data preparation phase began with the import of all required Python libraries. Three months of raw usage data were concatenated into a single DataFrame, while the master content reference (Master_Lookup) was loaded separately into its own DataFrame.

To improve performance and reduce processing costs, the initial step involved grouping the dataset by Content Name and Month, and aggregating total views accordingly. However, due to inconsistencies in the way identical programmes were recorded across different formats within the same month, these totals did not yet reflect true consumption. It became evident that additional normalization was necessary before meaningful insights could be extracted.

Parsing content names

Rather than applying fuzzy matching directly on long string combinations of content title, episode name, season, and episode numbers, a structured approach using regular expressions (regex) was adopted. This decision was made to maintain computational efficiency, particularly when handling large datasets.

A total of 9 parsing patterns were developed to extract the following key fields from the Content Name column: Programme Title, Episode Title, Season Number, Episode Number, Quality Label.

Each pattern was tailored to account for unique formats and exceptions. When exceptions were detected, manual corrections were applied. Although if-else blocks are typically preferred in production code for readability and maintainability, regex was used here for rapid prototyping, supported with extensive inline comments for clarity.

Matching with lookup

Once parsed, both datasets—the raw and the master—underwent standardization. This included converting all text to lowercase and removing multiple whitespace characters. Additional noise, such as .. used in place of unknown episode titles, was removed to improve the quality of fuzzy matching.

Matching was executed in two stages:

1. Fuzzy matching on programme titles between the parsed dataset and the master lookup.
2. For each matched programme, a second fuzzy match was applied on the episode title, limited to entries within that programme only—greatly enhancing accuracy and performance.

Thresholds were set based on manual tuning:

1. Programme match score ≥ 60%
2. Episode match score ≥ 80%

Only records exceeding these thresholds were retained for downstream analysis.

Two key output scenarios emerged:

1. Records with matched programme, episode title, season, and episode numbers
2. Records with matched programme, season, and episode numbers, but missing episode titles

For Scenario 1, the matched season and episode numbers from the master lookup were preserved. Scenario 2 could not be validated against the lookup and was therefore excluded from the join step.

Final dataset structure

The two cleaned subsets were recombined and grouped by: month, matched_programme_title, parsed_season, parsed_episode.

This aggregation produced the final viewership figures per episode per month. These results formed the basis for all further analytical insights.

3. Insights & Observations

Most watched programmes

The analysis of total view counts over a three-month period reveals a strong preference for documentary and mystery-driven content. Programmes such as “bradley walsh: egypt’s cosmic code”, “ancient aliens”, and “the curse of oak island” emerged as the top performers, each accumulating over 100,000 views.

Interestingly, these titles share common thematic elements—history, ancient civilizations, and speculative narratives—which suggests a content cluster that strongly resonates with Platform X’s audience. Moreover, the distribution of viewership across the top five indicates that users engage with a variety of titles rather than concentrating on a single trending show.

Key Takeaways:

1. Viewers show consistent engagement with intellectually stimulating and curiosity-driven genres.
2. Data can inform targeted content acquisition in similar thematic categories.
3. Promotional efforts might yield better ROI when focused on historically or scientifically themed shows.

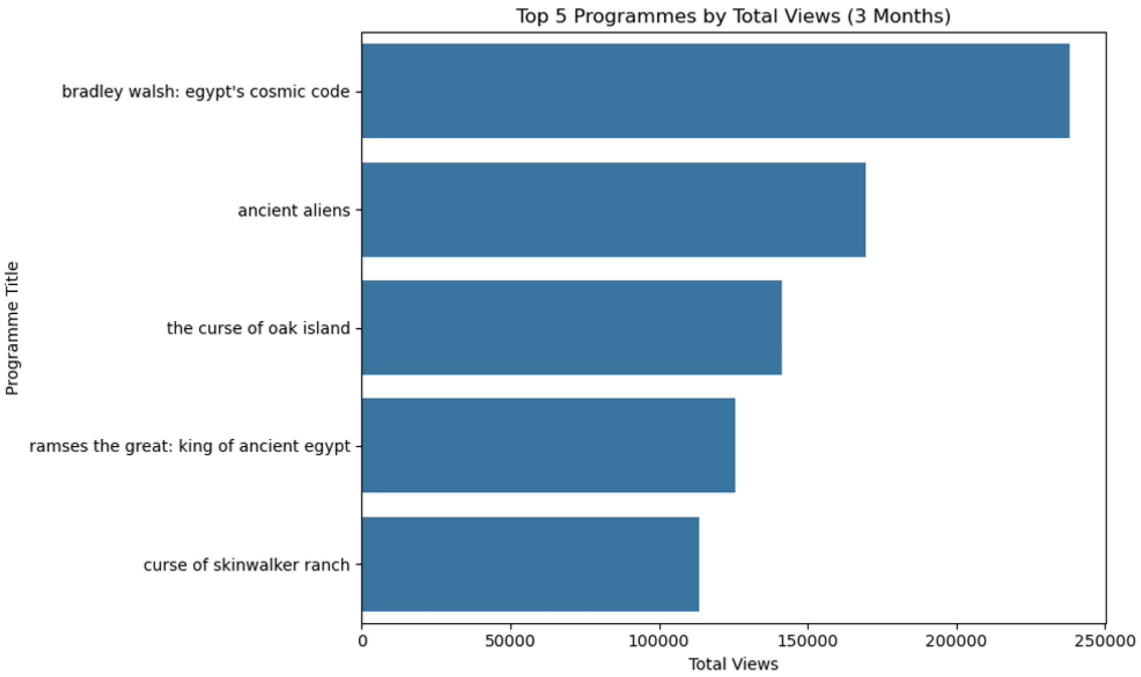

Figure 1: Top 5 Programmes by Total Views (3 Months)

Temporal Viewing Trends

Monthly aggregated view counts reveal a distinct fluctuation in audience engagement. The month of March shows a significant peak in viewership, exceeding 1.2 million total views, while April and May present a consistent decline, each stabilizing below one million.

This trend may suggest the impact of seasonal factors or content scheduling strategies on platform engagement. The sharp contrast between March and the following months indicates an opportunity to examine viewer behavior or content programming decisions during this timeframe.

Key Takeaways:

1. March experienced the highest viewership, indicating strong platform activity at the start of the quarter.
2. April and May maintained relatively stable but lower engagement levels.
3. Additional analysis could identify whether the March peak was driven by new content, marketing efforts, or seasonal demand.

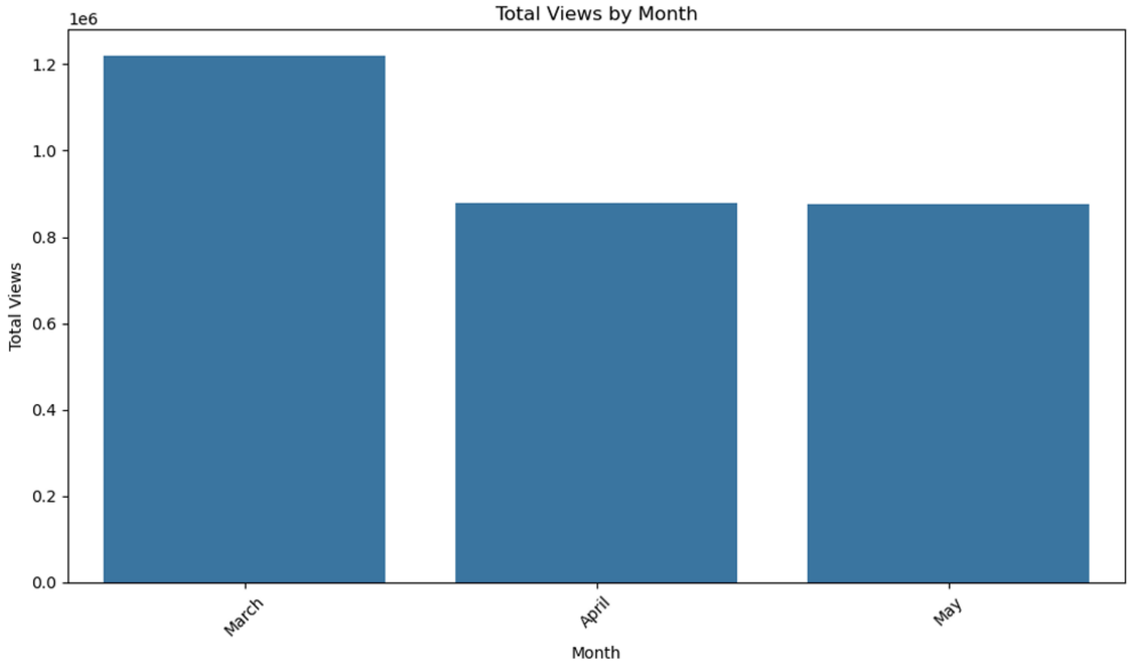

Figure 2: Total Views by Month

Viewer Retention Across Episodes

To gain deeper insights into user engagement over the course of a season, a drop-off analysis was performed for Season 1 of the most viewed programme, “bradley walsh: egypt’s cosmic code”.

Interestingly, the chart indicates a slightly increasing trend in viewership across the three episodes, with episode 3 receiving the highest number of views among them. This suggests that the content maintained or even gradually increased viewer interest over the short season.

Contrary to common expectations of a mid-season or final-episode drop in attention, the data indicates a stable or positively inclining viewership curve, hinting at strong retention or potentially increased interest driven by word-of-mouth or compelling storytelling.

Key Takeaways:

1. Viewership remained steady or increased slightly across episodes, showing no early drop-off.
2. The short season likely helped retain attention and reduce fatigue.
3. Such trends may support strategic decisions to keep documentary-style series concise and tightly written.

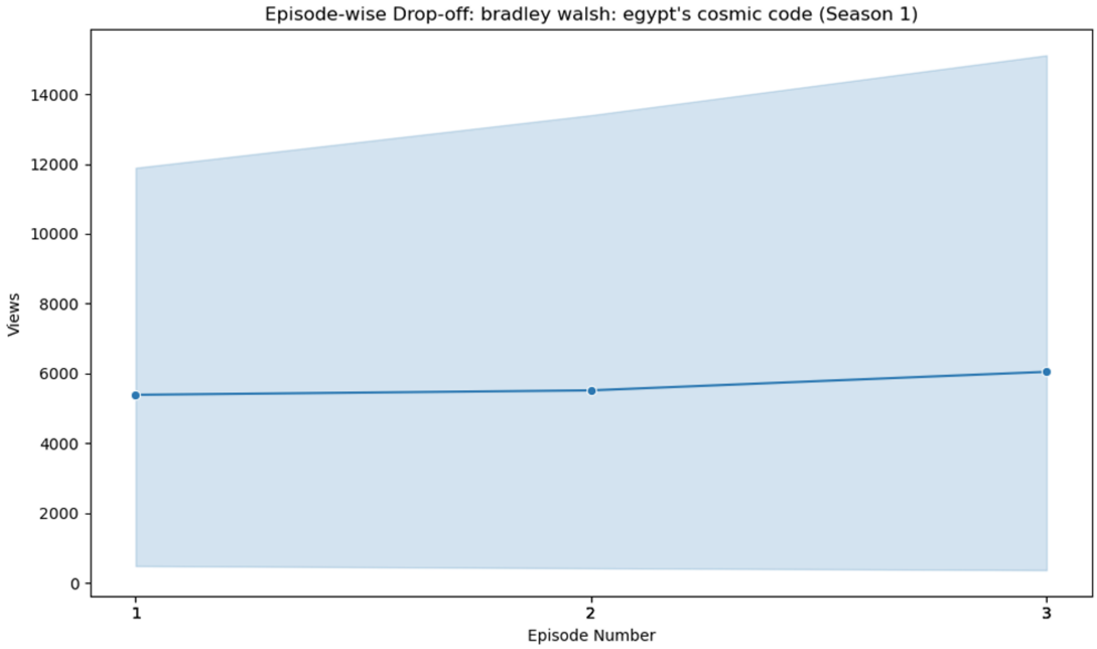

Figure 3: Episode-wise Drop-off

Monthly Performance Comparison of Top Programmes

To understand how audience interest evolved over time for the most popular programmes, a monthly breakdown of views was analysed for the top 5 titles across March, April, and May.

The chart reveals that “bradley walsh: egypt’s cosmic code” and “ramses the great: king of ancient egypt” experienced sharp drops in viewership after March. In contrast, "ancient aliens", "the curse of oak island", and "curse of skinwalker ranch" maintained relatively steady levels of audience engagement across the three months.

This divergence suggests that certain programmes had initial hype or promotional surges that drove high viewership in March, which was not sustained. Others appear to have more stable fan bases or benefitted from ongoing content appeal.

Key Takeaways:

1. Titles like bradley walsh saw a dramatic fall after peak initial interest, highlighting short-lived attention cycles.
2. Steady performers like ancient aliens suggest consistent viewer loyalty and long-term engagement potential.Marketing and scheduling strategies could benefit from identifying which content types drive immediate spikes versus sustained interest.

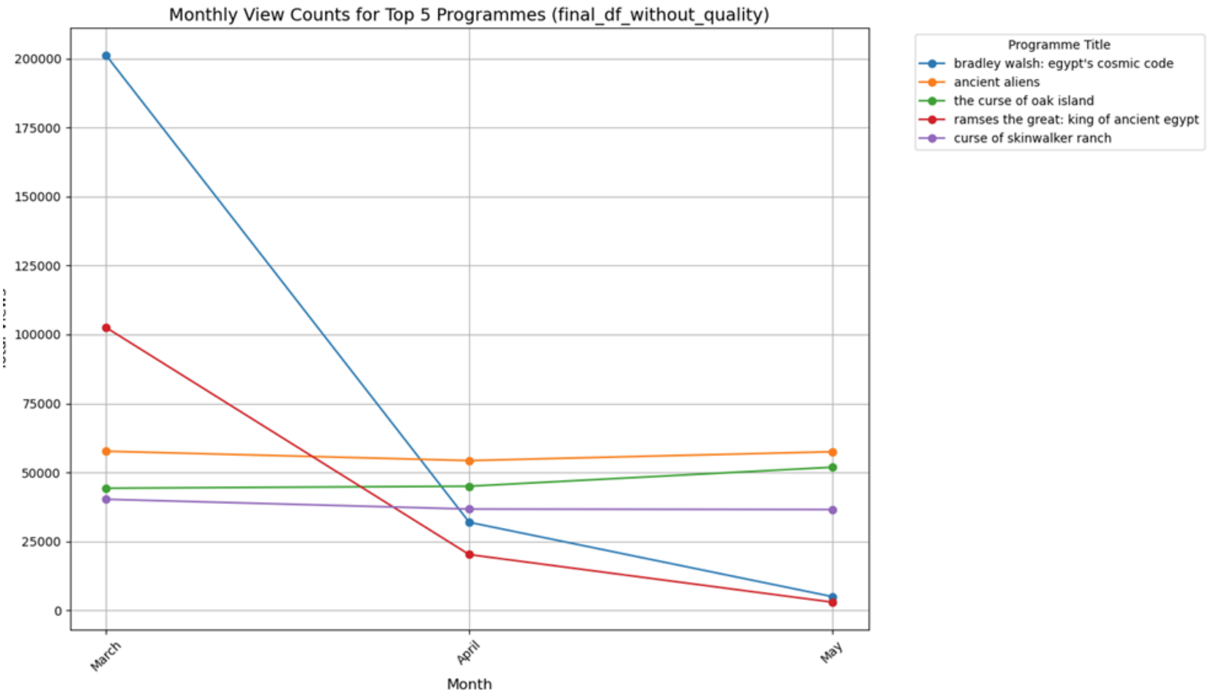

Figure 4: Monthly View Counts for Top 5 Programmes

Episode-Level Viewership Trends

To evaluate engagement trends across all programmes, average view counts were calculated per episode number. This helps reveal how viewer interest changes as episodes progress, regardless of the programme.

The graph demonstrates a noticeable drop in average views as episode numbers increase. The first few episodes receive significantly higher average views, which then taper off rapidly by episode 10–20, eventually stabilizing at much lower levels. This is a classic indication of early drop-off in episodic content, where audiences often sample initial episodes but fail to continue long-term.

There is also high variance in early episodes, suggesting inconsistent performance — possibly driven by mixed promotion, content quality, or scheduling. Beyond episode 50, average viewership stabilizes at a low level, with minimal fluctuations.

Key Takeaways:

1. Significant early drop-off in average views suggests challenges in retaining audience attention.
2. Episodes beyond the 20th tend to receive substantially fewer views.
3. High variance in early episodes may reflect inconsistent programme appeal or discoverability.
4. Findings support the need to optimize the first episodes for impact and explore strategies for better long-term retention.

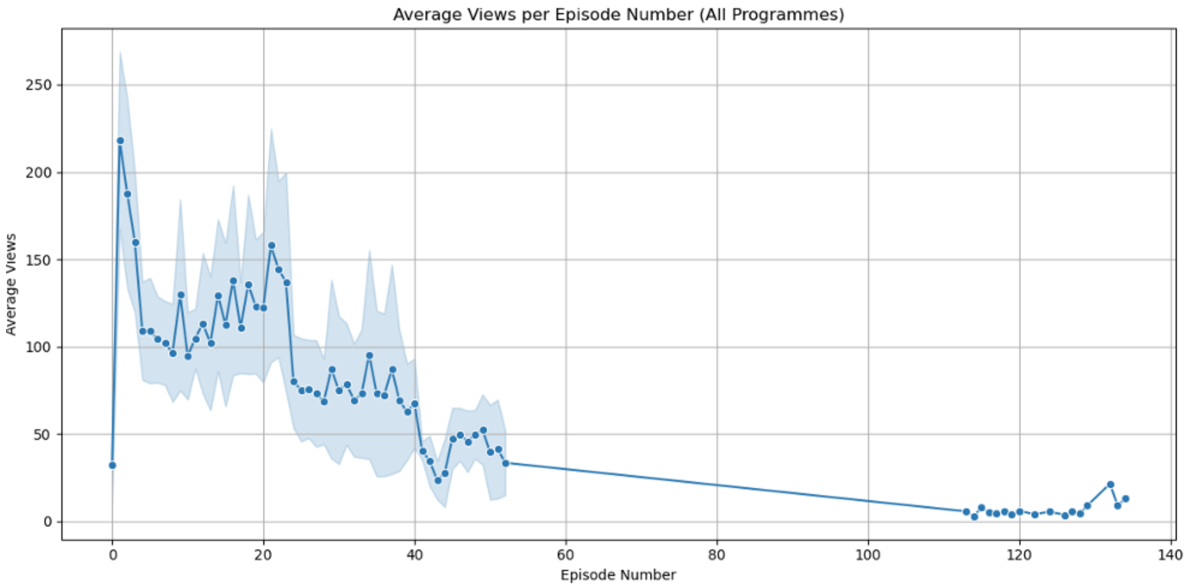

Figure 5: Average Views per Episode Number

Viewership Consistency

To understand how stable viewer engagement is throughout different episodes, a standard deviation analysis was conducted on episode-level viewership data. This identifies programmes where episode views vary the most, suggesting uneven audience retention or fluctuating interest levels.

The top 10 programmes with the highest standard deviation in view counts include "bradley walsh: egypt’s cosmic code", "ramses the great: king of ancient egypt", and "jack the ripper: written in blood". These titles exhibited the largest inconsistencies in viewership between episodes.

Such variability may result from inconsistent episode quality, viewer fatigue, or external promotional factors. It's also possible that some episodes address more engaging or trending topics than others.

Key Takeaways:

1. High standard deviation indicates volatile audience engagement across episodes.
2. Programmes with historical or crime themes appear frequently in the list.
3. Identifying fluctuation patterns helps in improving episode planning and promotional strategies.

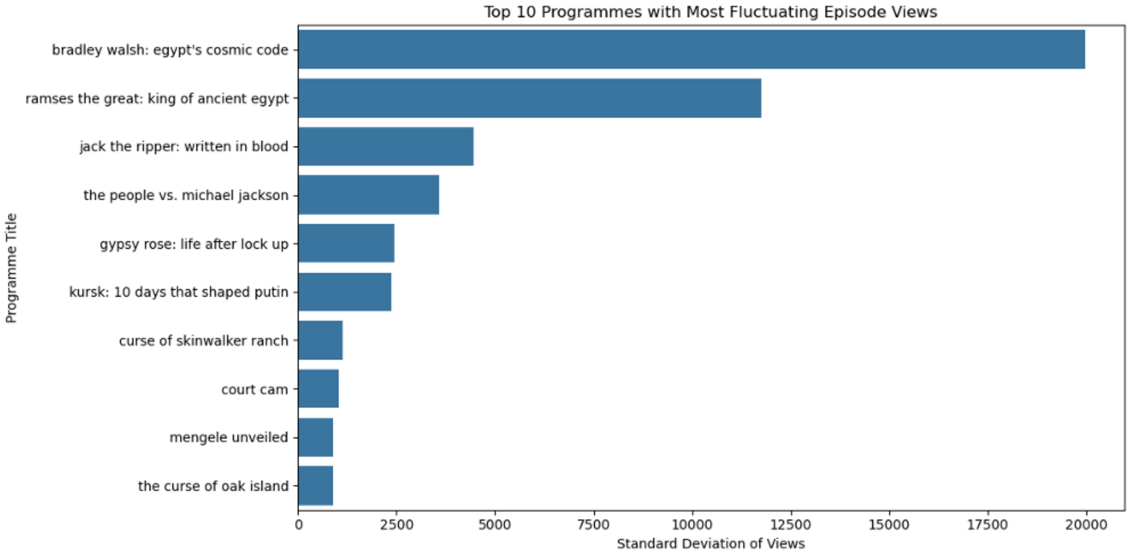

Figure 6: Top 10 Programmes with Most Fluctuating Episode Views

Longevity vs Audience Engagement

To evaluate how programme longevity relates to audience engagement, we examined the top 10 programmes by number of episodes and analysed them across three dimensions: total episode count, total views, and average views per episode.

The first bar chart highlights that “Killer in Plain Sight” leads in terms of episode volume, followed closely by “The UnXplained with William Shatner”, “Auction Kings”, and “Storage Wars”—all having between 70 and 100 episodes. These shows represent the longest-running content in the dataset.

However, the scatter plots reveal a more nuanced story:

1. In the Episode Count vs. Total Views chart, “Ancient Aliens” outperforms other long-running shows in terms of cumulative audience, despite not having the highest episode count. On the contrary, “Killer in Plain Sight”, although longest, shows moderate total views.

2. In the Episode Count vs. Average Views per Episode chart, “Hometown Tragedy” emerges as a standout, achieving the highest average views per episode among its peers—despite a shorter run. This indicates strong per-episode engagement and suggests high content impact in fewer episodes.

These results highlight that longevity does not necessarily translate to sustained high viewership. Some shorter or mid-length series may engage their audience more intensely on a per-episode basis, while longer series may struggle to maintain momentum.

Key Takeaways:

1. Longer series like “Killer in Plain Sight” and “Auction Kings” don't always secure top viewership or average engagement.
2. “Ancient Aliens” shows strong total viewership despite having fewer episodes than the longest-running programme.
3. “Hometown Tragedy” proves that a focused, shorter series can outperform longer ones in average viewer engagement.

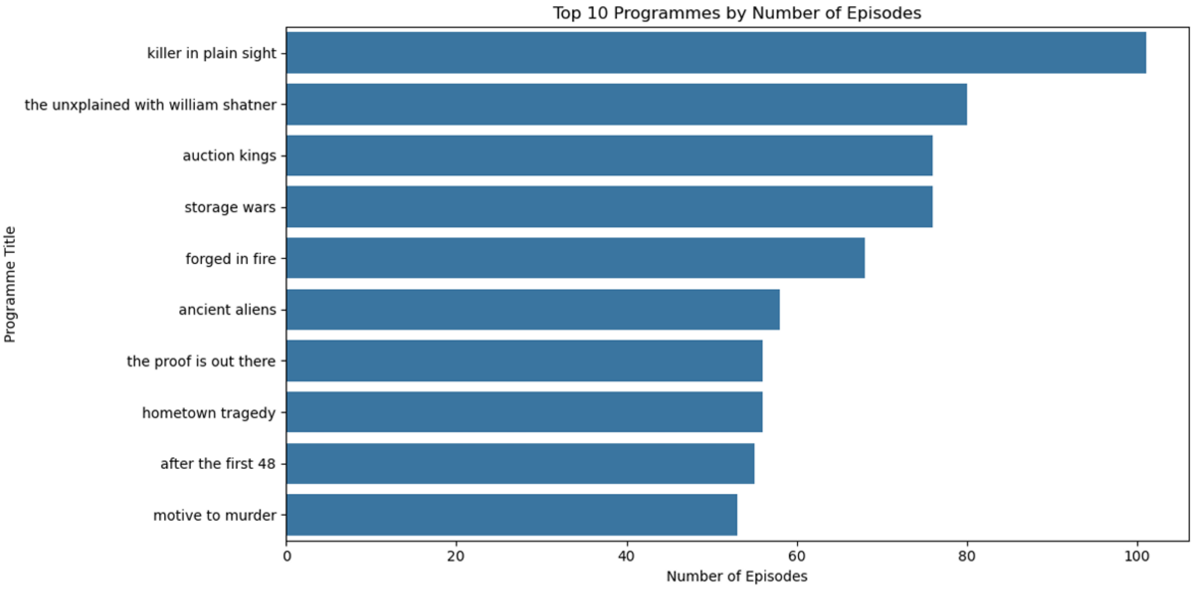

Figure 7: Top 10 Programmes by Number of Episodes

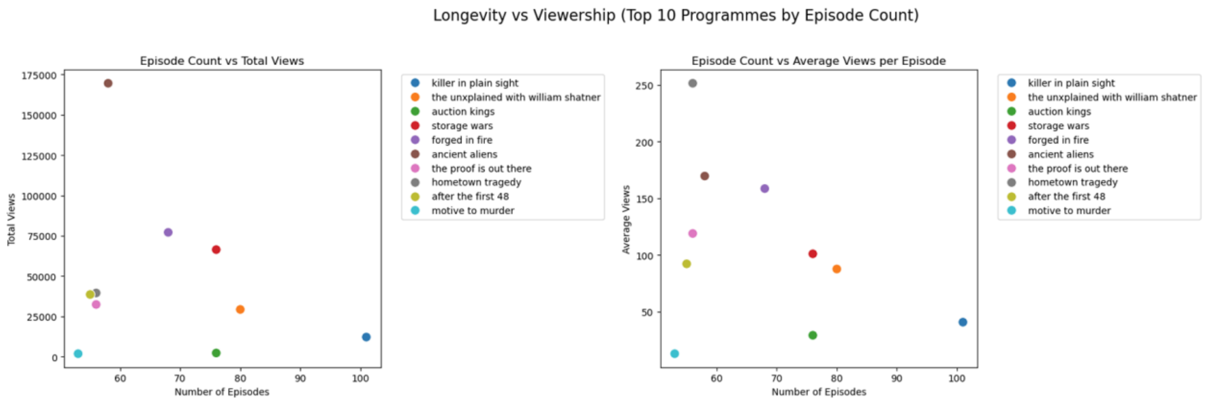

Figure 8: Longevity vs Viewership (Top 10 Programmes by Episode Count

4. Conclusion

This analysis reveals that episode longevity alone does not guarantee strong audience engagement. While some long-running programmes maintain consistent viewership, others—despite high episode counts—struggle to attract large audiences or sustain interest over time. Conversely, a few shorter series demonstrated significant impact through higher average views per episode, indicating a concentrated audience interest in more focused content.

In particular, "Ancient Aliens" proved successful in reaching a large cumulative audience, while "Hometown Tragedy" achieved the highest average viewership per episode—suggesting that quality and relevance may outweigh quantity in driving engagement.

These insights underscore the importance of evaluating both quantitative and qualitative dimensions when assessing programme performance. Broadcasters and content strategists should therefore consider not just episode count but also viewer behavior patterns to shape effective content development strategies.