# **Signals in the Scroll**

# ***— EDA Narrative Report***

This report starts from a single raw CSV, performs data cleaning and harmonization, and builds an interactive interface in Dash with a unified visualization grammar in Plotly. Because every figure follows the same display rules, readers can compare evidence and conclusions across questions with fewer confounds. To reduce cognitive load, we adopt a “minimal necessary encoding” strategy: color encodes platform only; percentage axes share a consistent tick format; key statistics (medians, slopes) are labeled inside the figure; annotations are anchored to the paper coordinate system to avoid drifting on zoom; and log–log plots use explicit base‑10 ticks for stable cross‑environment rendering. Together these choices form a framework that is readable, comparable, and reproducible.

## **Methods and Reproducibility**

Before addressing each question, we standardize data and methods. File discovery uses progressive fallback: prefer a list of candidate filenames, then scan by keywords, and finally fall back to any CSV so the notebook runs under different directory layouts. After loading, behavior‑related numeric columns are coerced with to\_numeric, replacing infinities. When missing, the helper ensure\_ratio reconstructs the three component rates (like/comment/share ÷ views) and the overall engagement rate (as their sum). Time variables are normalized: publish\_dayofweek is mapped to an ordered Mon→Sun category with is\_weekend derived; upload\_hour is constrained to 0–23. For large scatterplots we use within‑platform equal sampling to balance performance and representativeness. In log‑scale charts (e.g., views vs likes), slope estimation and plotting share the same positive‑only filtered subset to keep computation and presentation aligned. Color mapping and platform order are fixed to TikTok→YouTube throughout. All figures can be batch‑exported via Kaleido to PNG/SVG/HTML, with filenames following a “numeric prefix + title slug” convention for easy retrieval and ordering.

Under these methodological constraints, we first establish a platform baseline to avoid interpretation bias from mixing platforms. We then examine posting time, content duration, and topical framing, and finally turn to the supply side and trend persistence—closing the loop from user behavior to content and production.

## **Q1 Platform Differences: Establishing a Comparative Baseline**

We begin with platform‑level distributions (Figure 1). The boxplots show that TikTok exhibits a higher engagement‑rate distribution than YouTube; median labels above the boxes (≈8.9% vs 4.9%) provide a quotable quantitative reference. Pooling platforms risks a Simpson’s paradox and can understate the systematic effect of platform on engagement.

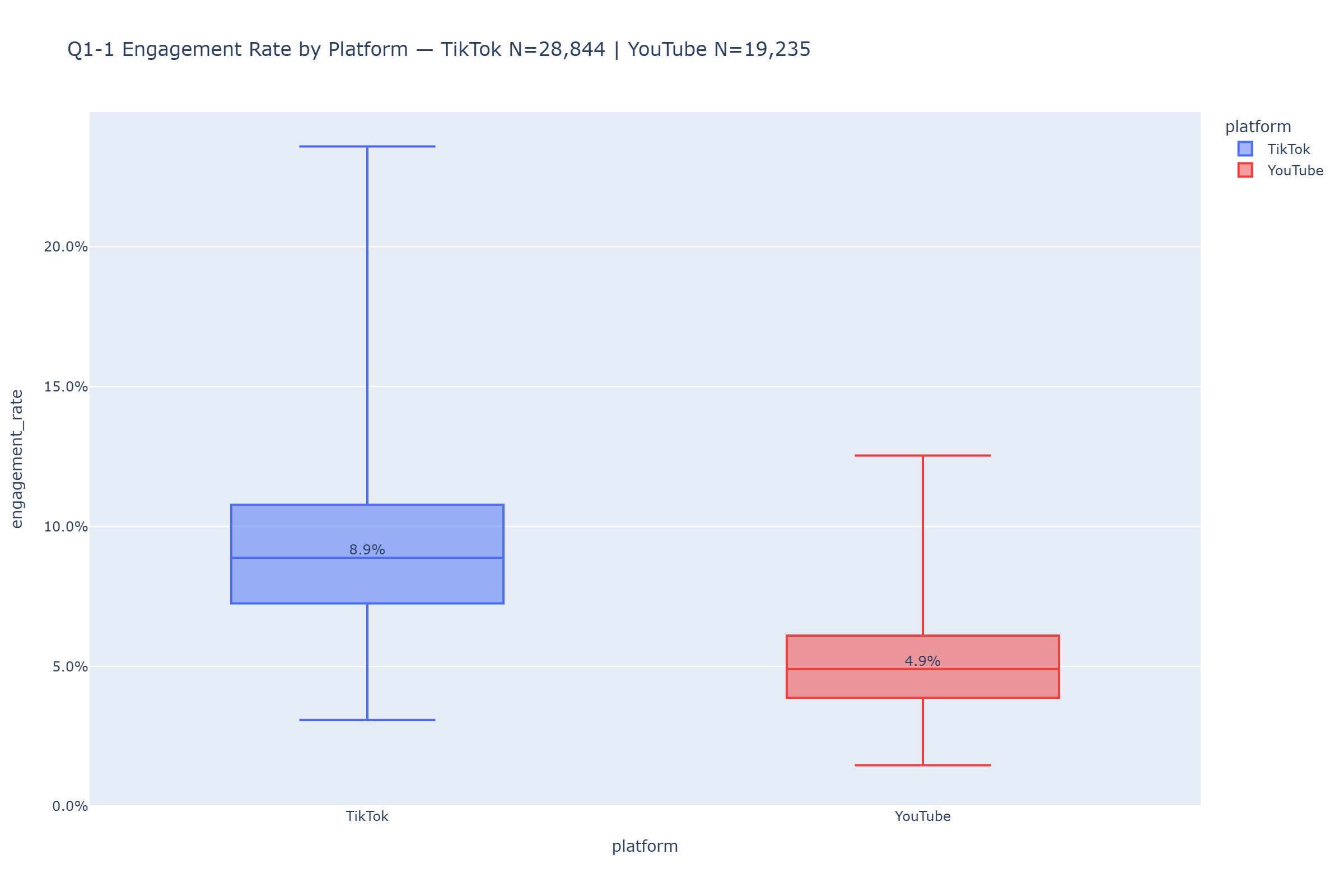


Figure 1 Engagement Rate by Platform

Building on this, we compare the structure of engagement (Figure 2). Converting like, comment, and share rates into long form and plotting them in the same coordinate system reveals consistently higher ratios on TikTok, especially for comments and shares. The difference is therefore not only in overall engagement but also in the composition of interaction; studies of diffusion and operational design should account for this structural gap explicitly.

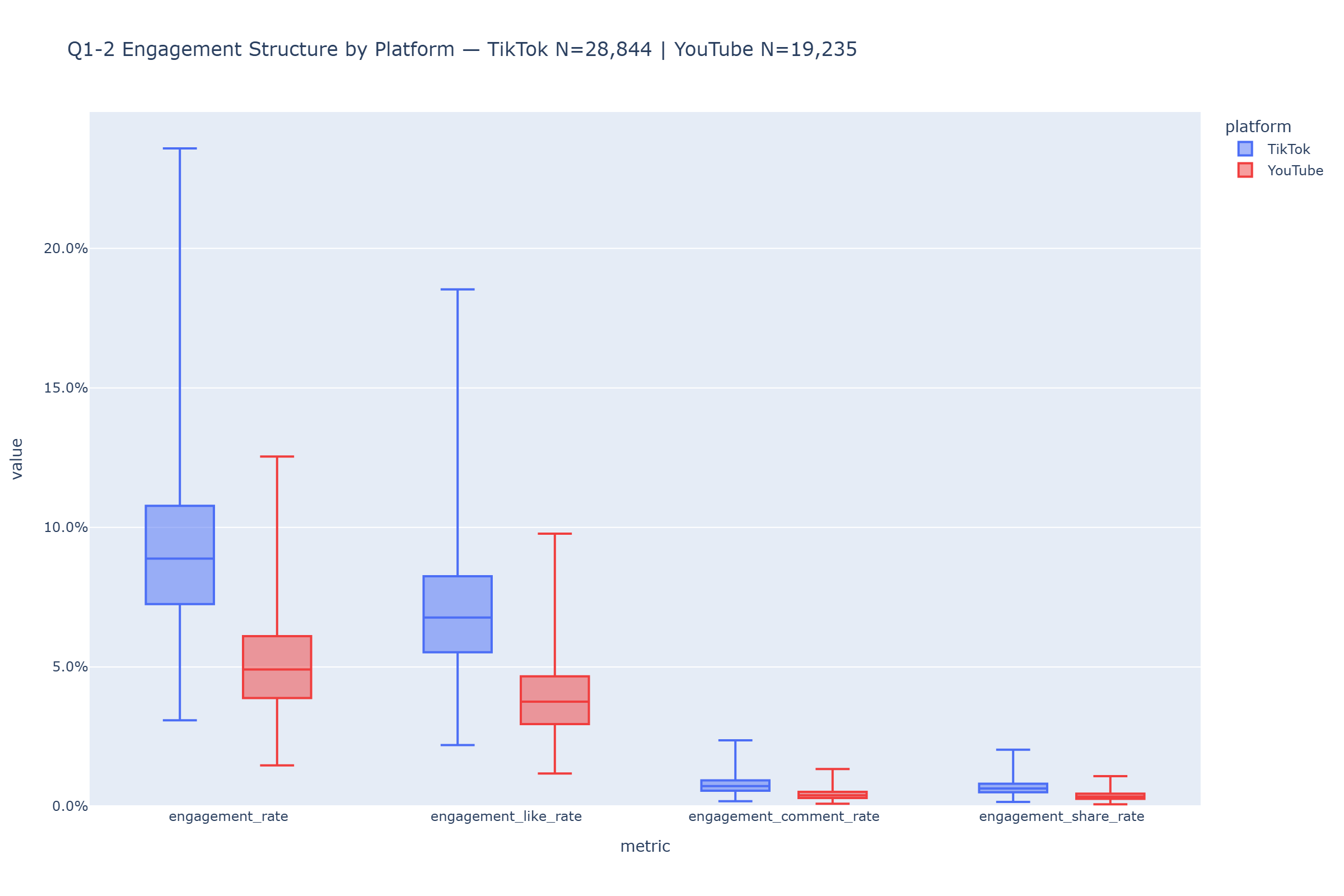


Figure 2 Engagement Structure by Platform

Next we examine how exposure translates into likes (Figure 3). In log–log space, plausible power‑law relationships become linear and slopes have a clear interpretation. TikTok’s point cloud lies higher overall and its regression slope is slightly larger (≈1.0022 vs 0.9993), indicating a more efficient conversion from views to likes at comparable exposure. Slope estimation and plotting use the same positive‑only subset, and the annotation is anchored in paper coordinates to avoid occlusion when zooming.

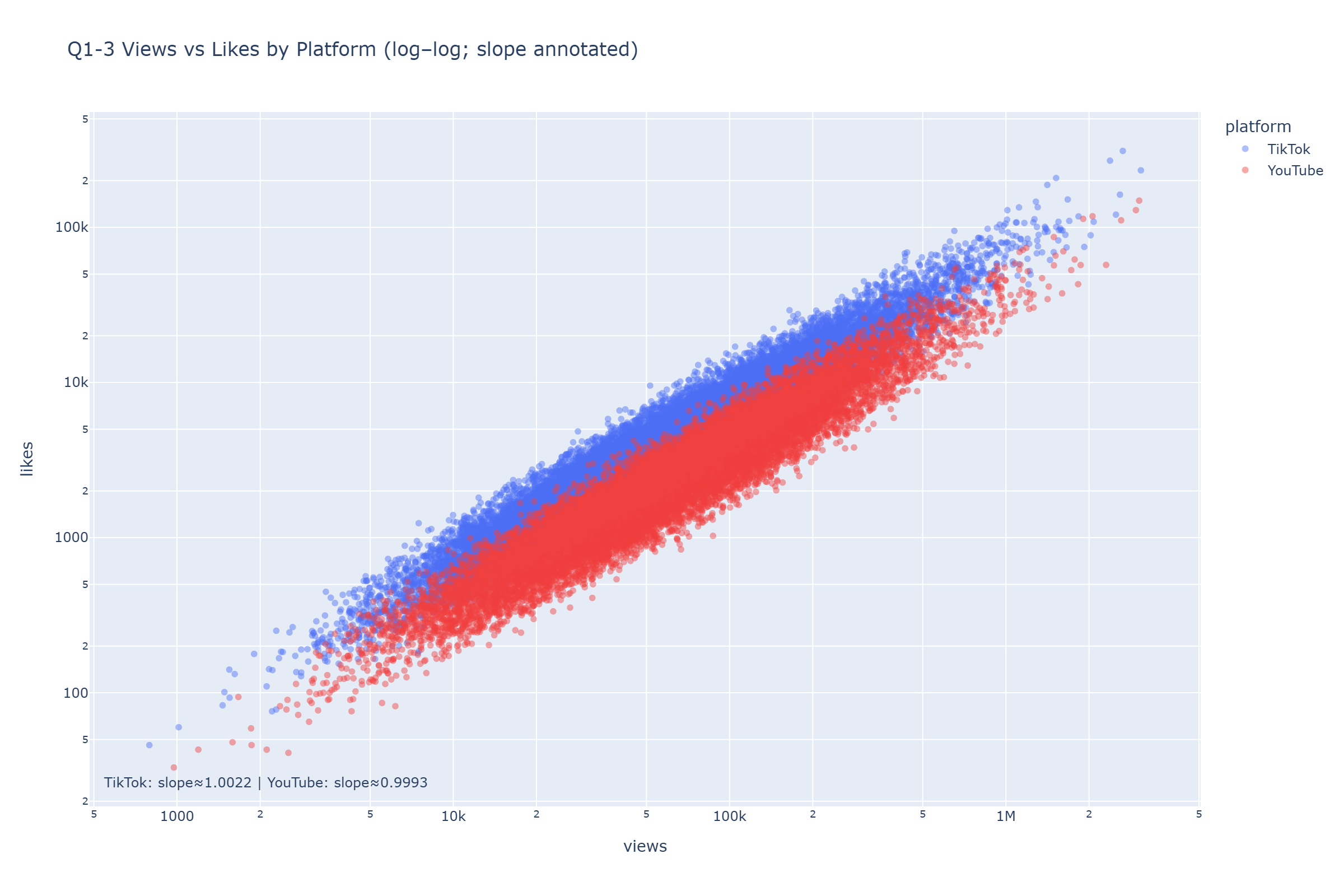


Figure 3 Views vs Likes by Platform

Once platform‑level differences are bounded, another direct driver of engagement is timing. We therefore shift to the temporal dimension—hours by weekday and the weekend–weekday contrast.

## **Q2 When to Post: Hour × Weekday and Weekend vs Weekday**

We first chart the full 7×24 grid (Figure 4). Median engagement across hours and weekdays reveals usable windows, yet patterns do not collapse into a single universal rule. A more defensible approach is to slice by platform and country in Dash and identify audience‑specific windows rather than apply a single “best time.”

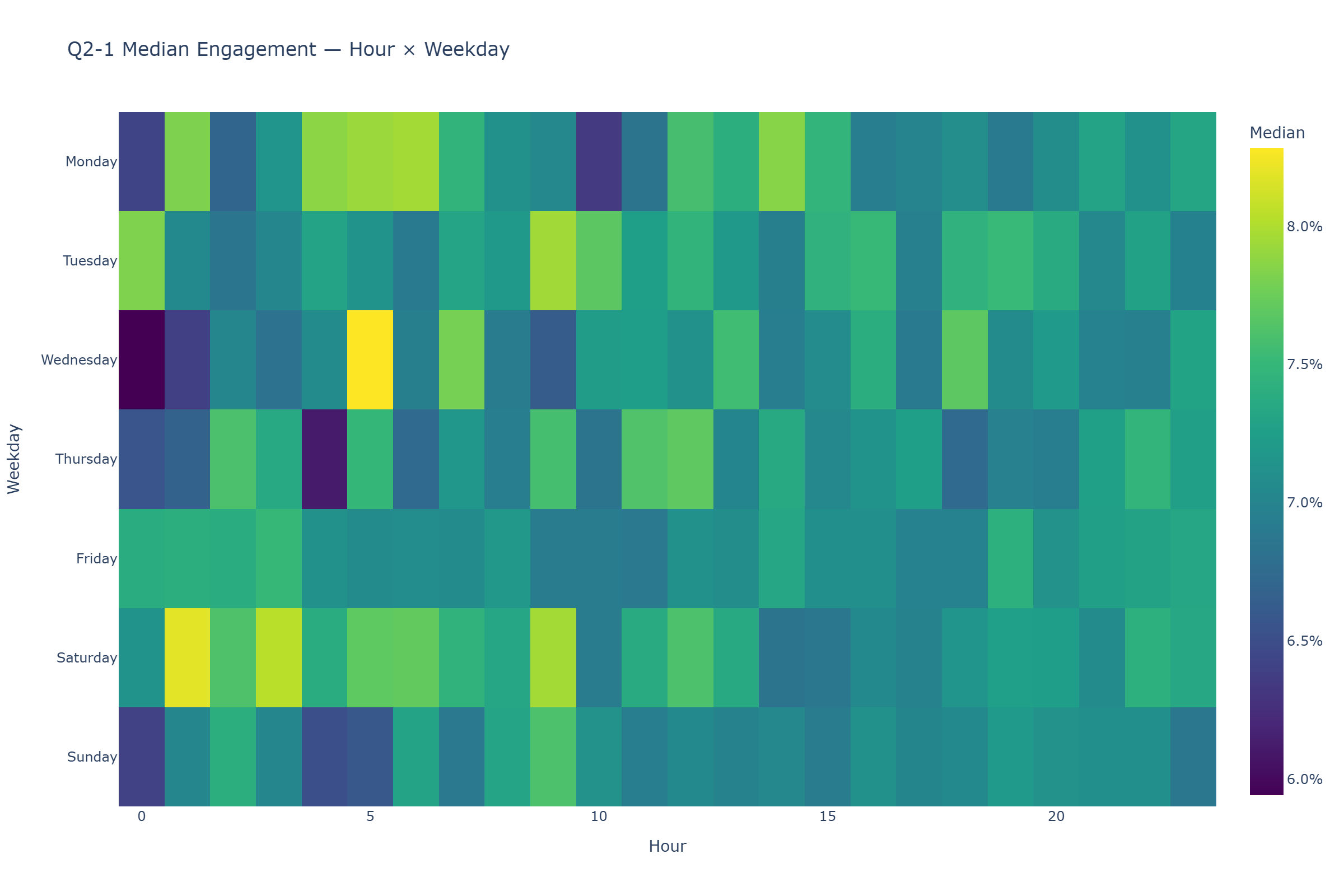


Figure 4 Median Engagement

We then test whether weekends are better (Figure 5). Two boxplots and the reported median difference show a small overall gap. This aligns with Figure 4: timing effects are context‑dependent, and specific strategies should be refined by platform and audience characteristics.

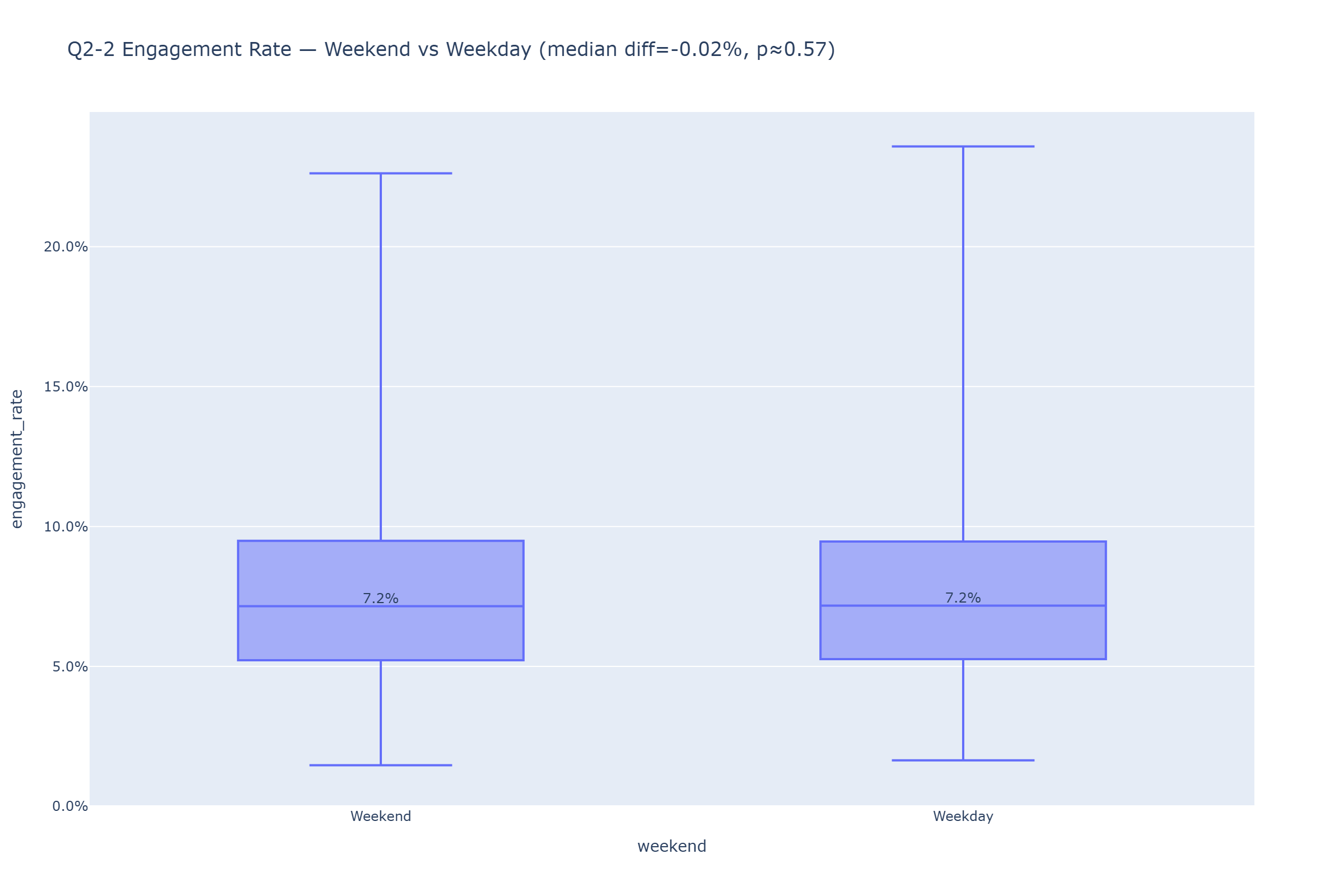


Figure 5 Engagement Rate Weekend vs Weekday

Timing answers “when to post,” but what is posted determines whether content is watched through. We therefore turn to duration and retention.

## **Q3 Duration and Retention: From Being Seen to Being Finished**

The duration × completion‑rate density (Figure 6) highlights a concentration zone: ≈20–35 seconds where completion is densest, with a marked decline beyond 60 seconds. This supports favoring the 20–35s range for stable completion; longer pieces need significantly higher information density and appeal in the first 3–5 seconds to counter retention pressure.

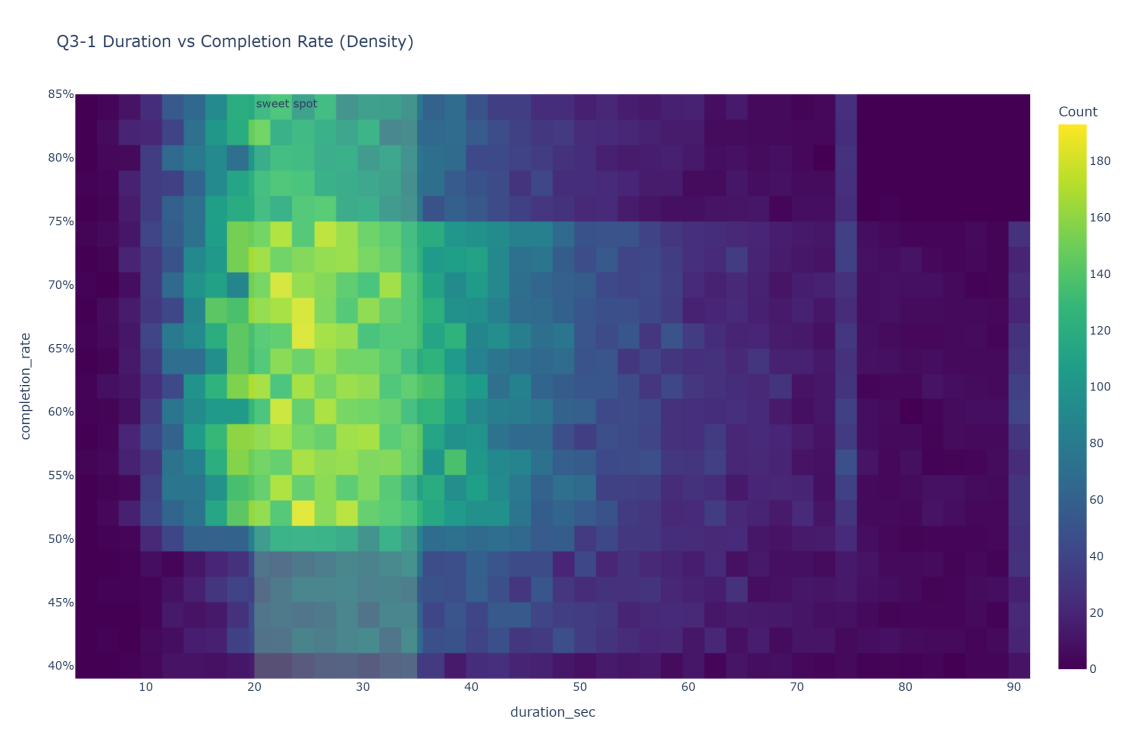


Figure 6 Duration vs Completion Rate Density

By platform, the relationship between duration and average watch time (Figure 7) rises with length on both platforms, but TikTok’s curve sits higher overall, implying better retention at equal durations. To preserve interactive performance, we use within‑platform equal subsampling and keep color encoding to platform only to reduce visual burden.



Figure 7 Duration vs Avg Watch Time by Platform

Having addressed “how long,” we next consider “what to talk about,” moving from categories to more granular topical tags.

## **Q4 Topics and Hashtags: From Categories to Finer Topical Units**

At the category level (Figure 8), median engagement for the Top‑20 categories (with sample sizes labeled) is sorted by median. Differences across categories are small, suggesting that choosing among broad categories alone is unlikely to deliver meaningful gains.

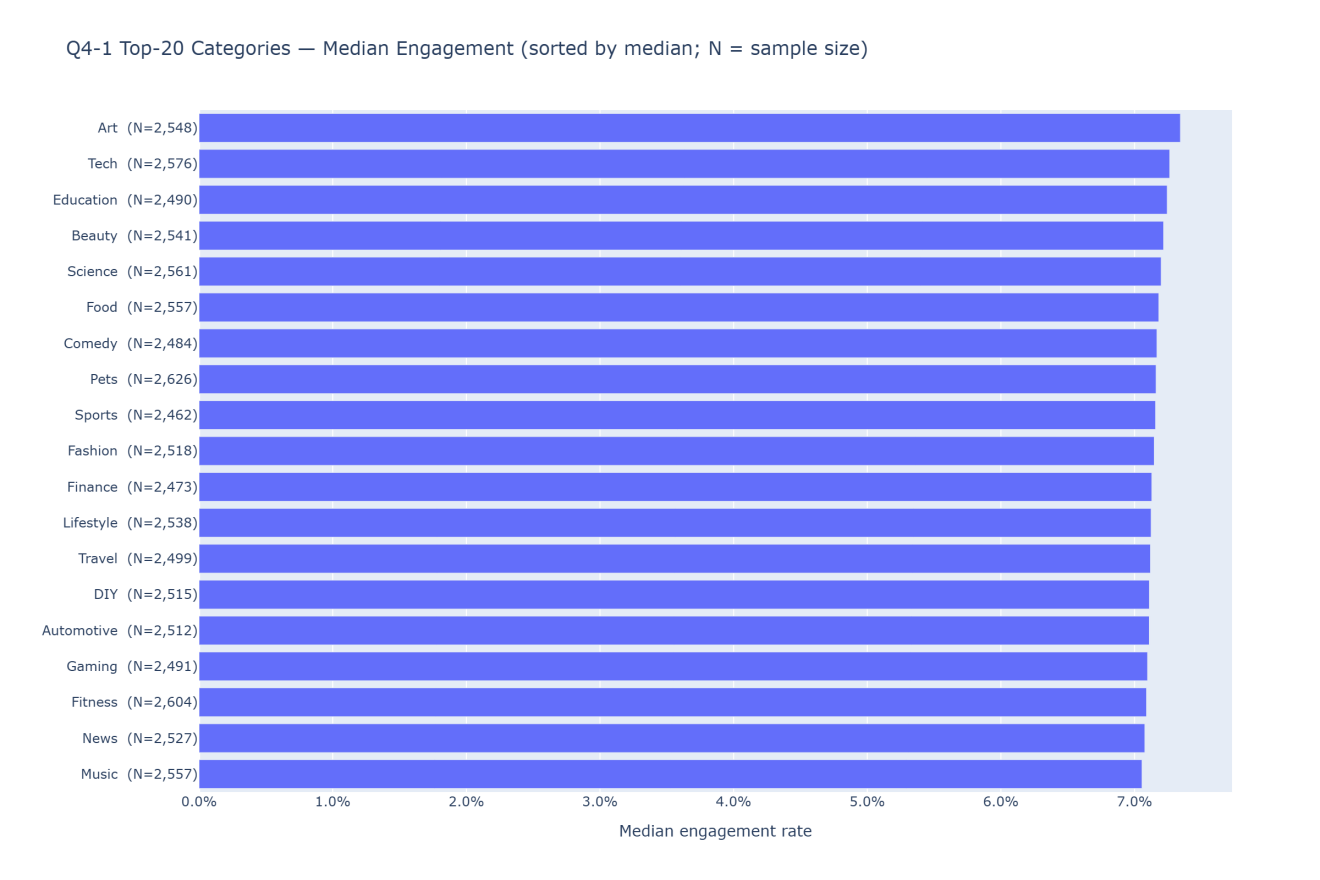


Figure 8 Top-20 Categories Engagement

Refining to the hashtag level (Figure 9), we compare median share rates for the Top‑30 topics. Overall differences are minimal—for example, the gap between Top‑1 and Top‑30 is about 0.04 percentage points—no apparent difference. Consistent with Figure 8, hashtag choice alone rarely produces substantive separation; a more prudent use is as a fine‑tuning variable within a given plan, validated via small A/B tests rather than treated as a primary optimization lever.

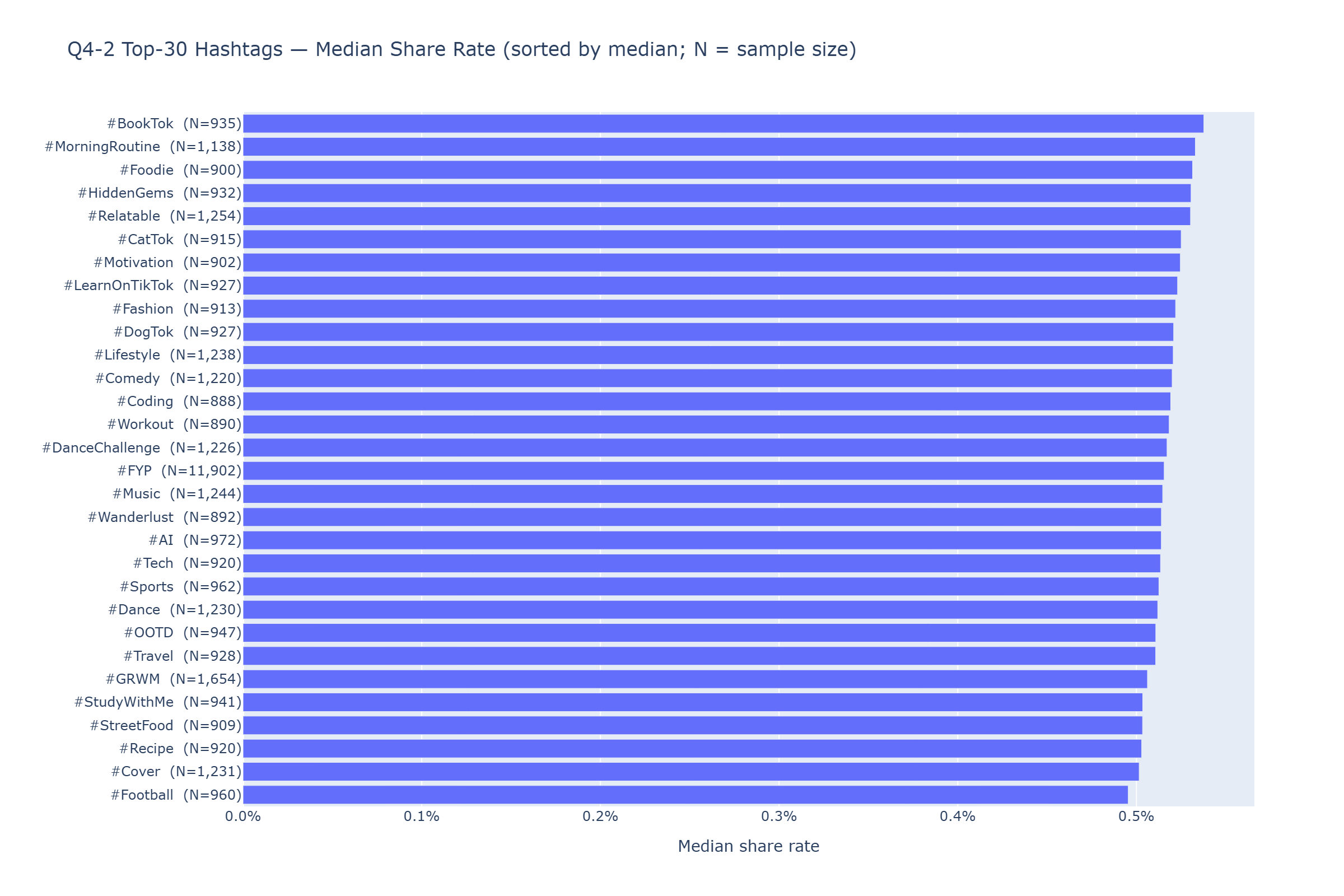


Figure 9 Top-30 Hashtags Share Rate

Topics answer “what to say,” but outcomes also depend on “who produces” and “how long momentum lasts.” We therefore return to the supply side and the trend lifecycle.

## **Q5 Supply and Lifecycle: Creator Tiers and Trend Persistence**

Comparing creator tiers (Figure 10) shows that mid‑tier creators have higher engagement distributions than micro‑tier creators (higher medians and upper quartiles). Scale, in other words, is not achieved at the expense of engagement, and mid‑tier partners can be prioritized in selection and placement.

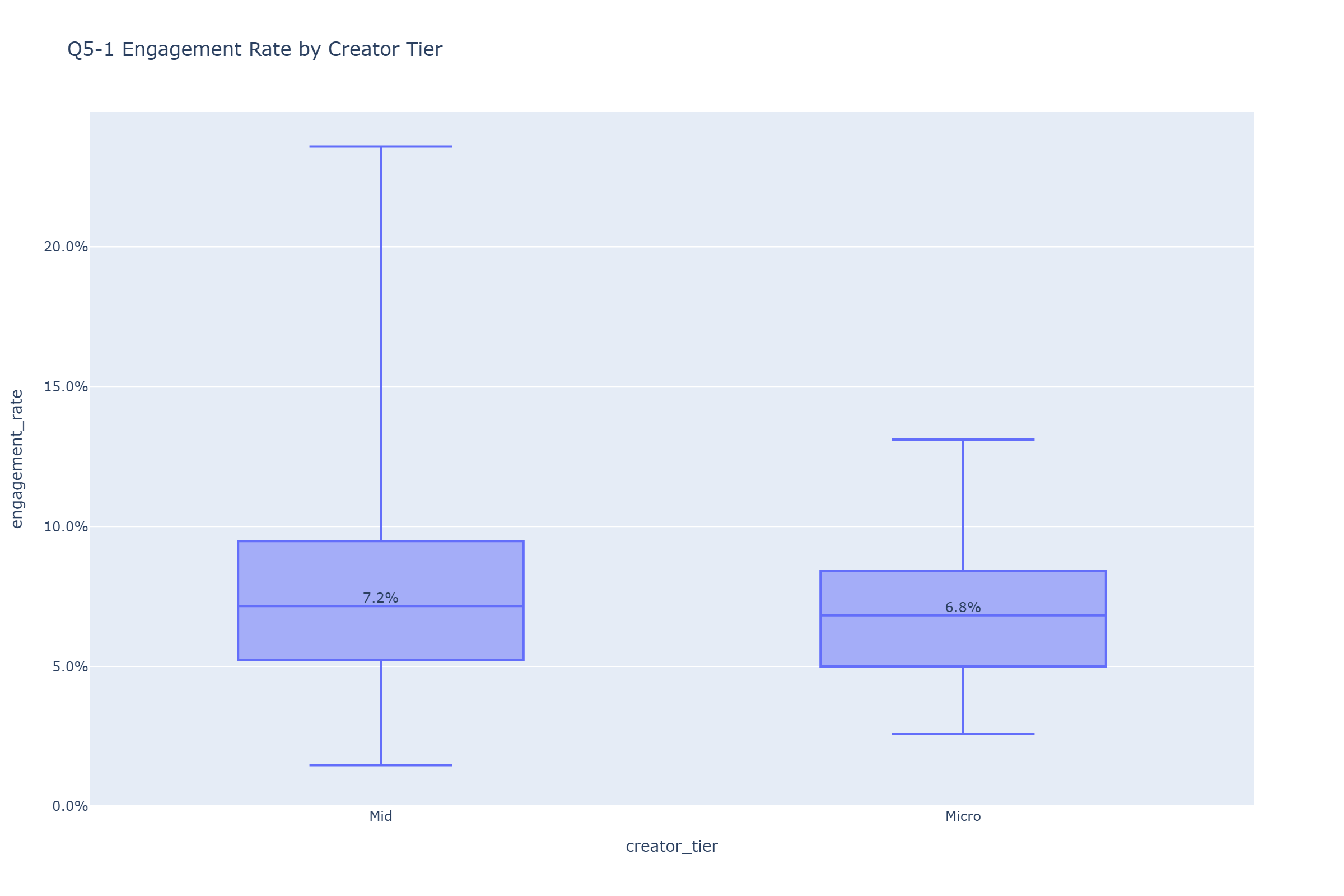


Figure 10 Engagement Rate by Creator

Trend‑lifecycle analysis (Figure 11) uses a log y‑axis to relate engagement velocity to trend duration, with bubble size encoding views. Most content peaks within 1–7 days and then declines; a minority combines high velocity with longer duration and warrants case‑by‑case review. Operationally, the first 3–7 days should be treated as a key observation window, with an “early acceleration” rule (allocate additional resources once velocity crosses a threshold), while curating a library of long‑lived, high‑velocity exemplars to guide future ideation and production.

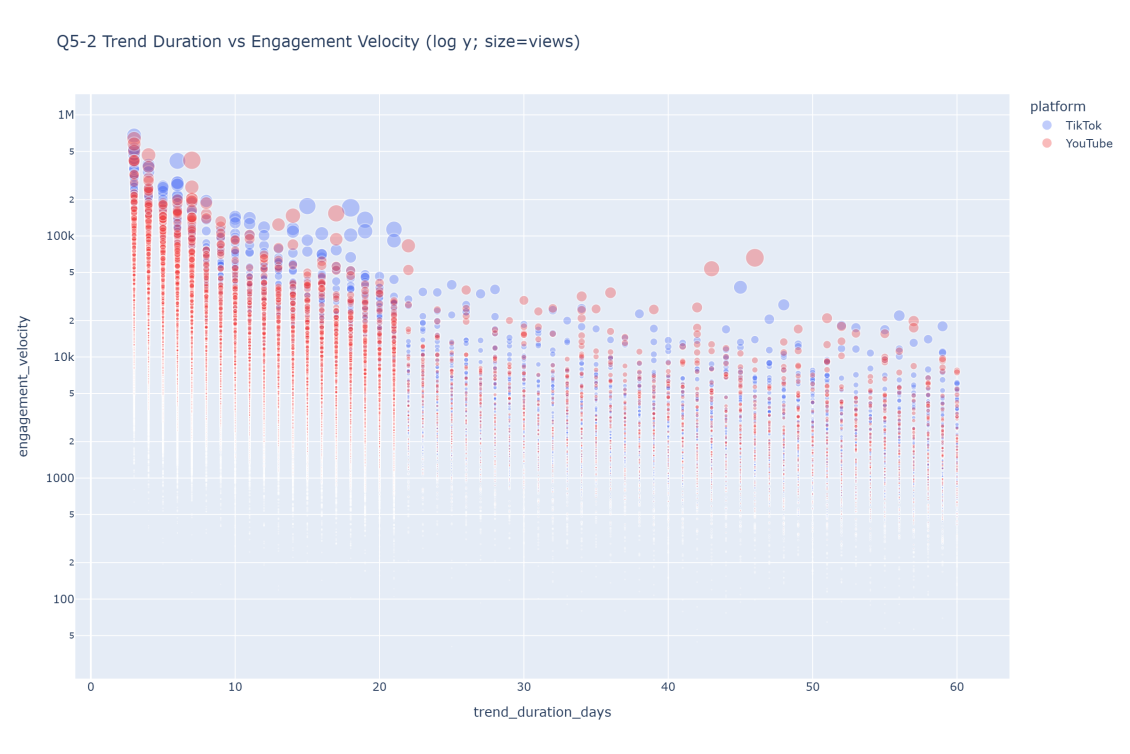


Figure 11 Trend Duration vs Engagement Velocity

The preceding sections specify what to do and how to do it. To present these findings reliably in an interactive setting, we also calibrate engineering details.

## **Engineering Calibration: Consistent Presentation in Interaction**

We standardize percentage formats, chart margins, and title templates; fix KPI‑card centering and decorative line alignment via CSS; specify base‑10 ticks for log–log plots and anchor slope annotations to paper coordinates; and apply within‑platform equal subsampling for scatterplots to balance performance and representativeness. These adjustments preserve readability and consistency across interactive views.

Even within a unified framework, it is useful to state analytical boundaries and possible extensions.

## **Limitations and Next Steps**

The current analysis relies primarily on univariate and bivariate comparisons and does not yet control for potential confounds such as country, category, and creator tier. Next steps include adding hierarchical or quantile regression behind Dash to test interactions like platform × duration × country; modeling non‑linear thresholds in the duration–retention relationship using splines or generalized additive models (GAMs); and reframing trend‑velocity evolution within an event‑history/survival‑analysis setup to quantify comparable half‑life metrics. On the reporting side, routinely display effect sizes (e.g., Cliff’s delta) and connect category/hashtag choices with an A/B testing framework to shorten the path from findings to deployment.

## **Conclusion**

Platform choice is the primary determinant of engagement performance. Posting windows and the duration–retention relationship provide additional levers. Differences at the category and hashtag levels are generally limited; hashtags are better treated as fine‑tuning variables and validated via A/B tests for marginal effects. On the supply and lifecycle dimensions, mid‑tier creators have a relative advantage, and the early window (days 3–7) deserves focused monitoring and resourcing. These conclusions are supported under a unified visualization and interaction scheme and can be embedded directly into operational decisions and experimental design.