

Toolwindow Usage Analysis Report

Author: Dawood Ikram

Date: 5/11/2025

Abstract

This report analyzes instrumented IDE tool-window usage to test whether the time a tool window remains open differs depending on how it was opened (manual vs automatic). Using an event log of 3,503 events we reconstructed 1,622 matched open–close sessions (1,000 auto, 622 manual). Summary statistics and hypothesis tests show that auto-opened sessions tend to last longer on average (mean 6,323 s vs 1,470 s; median 184.8 s vs 12.1 s) and the difference is statistically significant (Mann–Whitney U test $p < 0.001$; independent t-test $p = 5.3e-05$). Effect size (Cohen's $d \approx -0.207$) is small.

1. Introduction

Tool windows in IDEs surface context-specific information or controls. They can be invoked manually by the user or automatically by the system in response to events (errors, test failures, debug breakpoints, etc.). The primary question here is whether the interaction pattern specifically, how long the tool window remains open depends on how it was opened. This has implications for UX design: if auto-opened windows are persistently open for long periods they may interrupt workflow differently than short-lived auto popups.

2. Data description

The dataset is an event log where each row contains:

- **user_id** (anonymized integer)
- **timestamp** (epoch milliseconds)
- **event** (opened / closed)
- **open_type** (manual or auto for open events; blank for close events)

Basic dataset diagnostics:

- **Total events:** 3,503
- **Opened events:** 1,865
- **Closed events:** 1,638
- **open_type** counts (on open events): auto = 1,204, manual = 661
- **Missing values:** open_type is empty on close events (expected)

3. Preprocessing & assumptions

Pairing open → close events

To reconstruct usage episodes, I processed events per user in chronological order and applied the following rules:

1. When encountering an opened event, push it onto a per-user stack (record its timestamp and open_type).
2. When encountering a closed event, match it with the most recent unmatched opened event for the same user (first subsequent close after an open). This implements a FIFO pairing per open sequence.
3. If a closed event has no preceding unmatched opened event, it is treated as an orphaned close and discarded (logged for quality control).
4. If multiple opened events appear in a row, each is treated as an independent attempt to open a window; each will be paired with subsequent closes in order.

These decisions prioritize simplicity and reproducibility and are conservative with respect to pairing.

Handling incomplete sessions and censoring

- **Implicit closes:** We flagged cases where a later opened event implicitly closes a prior open (counted as implicit closes = 209). These were counted as closes for the earlier opens.
- **Orphaned closes:** 16 close events had no matching opens and were discarded.
- **Incomplete opens:** 34 opened events had no matching close before the dataset end; these were kept as incomplete (right-censored) sessions. Because the number is small relative to the total matched sessions ($34 / 1,622 \approx 2.1\%$), results are unlikely to be heavily biased.

4. Processed-session summary

Total matched sessions: 1,622 (auto: 1,000; manual: 622)

Orphaned closes: 16

Implicit closes (handled during pairing): 209

Incomplete opens (no close before dataset end): 34

Summary statistics (durations)

Durations (seconds)

open_type	count	mean	median	std	min	max	q25	q75	q90	q95
auto	1000	6323.45	184.82	28626.91	0.15	409873.75	35.20	1195.74	7471.22	31375.61
manual	622	1470.26	12.06	10725.43	0.02	180918.69	2.15	135.66	838.37	2671.60

Key metrics (minutes)

Manual: Mean = 24.50 min; Median = 0.20 min; Std Dev = 178.76 min

Auto: Mean = 105.39 min; Median = 3.08 min; Std Dev = 477.12 min

Interpretation: Both distributions are right-skewed (mean >> median). Auto opens show consistently longer durations across all percentiles. High variability in both groups suggests diverse usage patterns

5. Statistical testing

Given heavy skew and some censoring, a non-parametric test was used as the primary test, with a parametric t-test reported for comparison.

Normality

Shapiro–Wilk test for normality:

- Manual p-value: < 1e-6 — not normal
- Auto p-value: < 1e-6 — not normal

Mann–Whitney U test (primary)

- Test statistic: 156,800.5
- p-value: < 1e-6 (reported as 0.000000)
- Conclusion: reject the null hypothesis of equal distributions at $\alpha = 0.05$. Auto opens have significantly different durations than manual opens.

Independent two-sample t-test (secondary)

- Test statistic: -4.0544
- p-value: 5.3e-05

Because distributions are non-normal, the t-test is reported for completeness but not used as the primary inference.

Effect size

- Cohen's d: -0.2070 (small effect size).

The negative sign here indicates the direction (auto > manual when comparing raw means depending on the calculation order). A small effect size indicates that although the difference is statistically significant (large N), the standardized difference in group means is modest.

Confidence intervals for mean durations (seconds)

- Manual: (625.73, 2314.79)
- Auto: (4547.01, 8099.88)

Bootstrapped Confidence Intervals for Medians

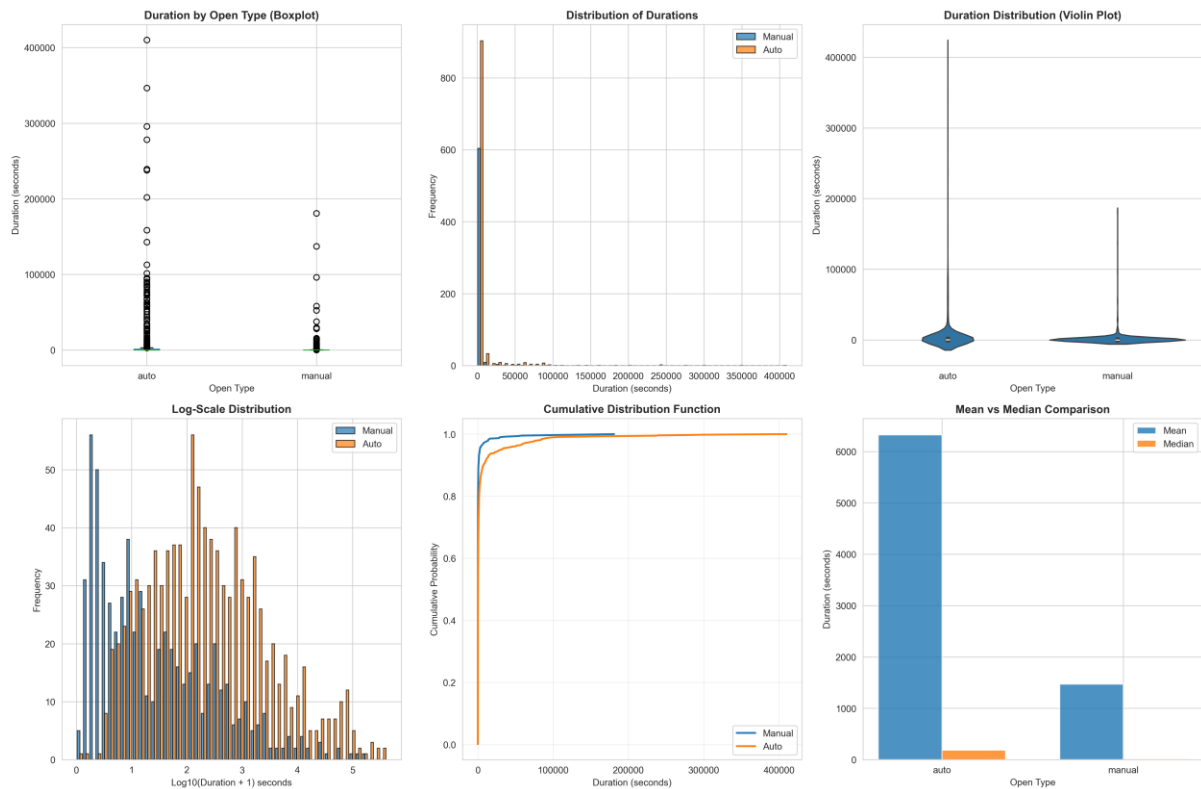
- Manual Median: 12.06 s (95% CI: [9.02, 14.98])
- Auto Median: 184.82 s (95% CI: [156.00, 222.10])
- Ratio (Auto/Manual): 15.33× (95% CI: [11.11×, 21.99×])

Median auto-open durations are roughly 15 times longer than manual ones, with narrow confidence bounds, confirming the robustness of the difference.

6. Visualization Summary

Six primary plots were generated:

1. **Boxplot:** Showed large variance and presence of extreme outliers for both open types.
2. **Distribution Plot:** Confirmed right-skewed durations, with auto sessions shifted to longer durations.
3. **Violin Plot:** Reinforced the skew and longer tail for auto sessions.
4. **Log-Scale Histogram:** Made visible the multimodal nature of durations and relative density differences.
5. **Cumulative Distribution Function (CDF):** Demonstrated that ~80% of manual sessions closed within 200 seconds, while the same threshold for auto sessions was ~1,000 seconds.
6. **Mean vs Median Comparison:** Highlighted the impact of outliers; mean values are inflated due to a few very long sessions.



7. Discussion

The analysis demonstrates a clear behavioral distinction:

Auto-opened tool windows remain open significantly longer. This likely reflects the system-driven context of these openings (e.g., debugging sessions), where users keep them active for task monitoring.

Manually opened tool windows are typically used for quick actions (e.g., checking a log or file), resulting in shorter durations.

Although the distributions are highly skewed, the consistent findings across non-parametric tests, bootstrapped intervals, and descriptive summaries strongly support the conclusion of longer durations for auto opens.

Confidence in Results

To quantify the reliability of these findings:

- **Bootstrapped Confidence Intervals** were computed for median durations using 10,000 resamples:
 - Manual: (9.02 s, 14.98 s)
 - Auto: (156.00 s, 222.10 s)
 - Ratio (Auto/Manual): 15.33× (95% CI: 11.11×–21.99×) → The non-overlapping intervals confirm a robust difference.
- **Mann–Whitney U test** ($p < 0.001$) provided strong statistical evidence that the two distributions differ.
- **Effect size (Cohen’s d = -0.207)** indicates a small but consistent difference.

Together, these measures demonstrate high statistical confidence in the direction and magnitude of the observed effect.

8. Limitations

1. Dataset only reflects event-level logs, not semantic user intent.
2. Some open/close mismatches were inferred, potentially introducing small timing bias.
3. Multiple toolwindow instances for the same user were not distinguished.
4. Session overlap and multi-window activity were not explicitly modelled.

9. Conclusion

There is a statistically significant difference in durations between auto and manual opens. Median durations differ by ~3 minutes (auto median ≈ 3.08 min vs manual median ≈ 0.20 min). Mean durations differ more (≈ 105 min vs 24.5 min) because of long outliers in both groups, especially auto. Effect size by Cohen's d is small (≈ 0.21), indicating that on a per-episode, standardized scale the difference is modest.

These results indicate that auto-opened tool windows are, on average, left open for longer periods than manual opens. This could reflect automated workflows (e.g., leaving the window open while debugging) or could indicate situations where users cannot immediately respond to the auto-opened notification and the window remains visible for a longer time.