Analysis Report

Analytics for Data Products IDEs

Task #1 Analyze Toolwindow Usage Data.

1. Approach

The primary objective was to determine whether the duration for which the tool window remains open differs significantly between manual and auto open types. To achieve this, the following steps were taken:

a. Data Cleaning & Parsing
- The dataset was loaded into a pandas DataFrame.
- Timestamp values were converted from epoch milliseconds to datetime.
- Event names were standardized to ensure consistency (e.g., "opened" → "open").

b. Event Matching Logic
- Events were grouped by user_id to ensure open/close pairs were matched per user.
- Each user's events were sorted by timestamp.
- A stack-based approach was used for pairing: opens were pushed onto a stack, and closes popped the most recent unmatched open.
- This handled cases like multiple opens in a row or closes without preceding opens.

c. Duration Calculation
- For each valid open/close pair, the duration (in seconds) was computed as:

$$\text{duration} = \text{close\_time} - \text{open\_time}$$

d. Statistical Comparison
- Summary statistics (mean, median, std) were computed for manual and auto opens.
- A two-sample t-test determined if mean durations differed significantly.

2. Assumptions

- Events are chronological within each user.
- Multiple open events before a close: only the latest unmatched open is used.
- Orphaned events (opens without closes, closes without opens) are ignored.
- Each user acts independently; durations are aggregated across users.

3. Handling Messy Data

a. Real-world event logs often include inconsistencies:
- Orphaned closes: ignored if no prior open.
- Multiple opens before a close: only the latest open is paired.
- Unclosed opens: ignored if no matching close before dataset end.

Only valid open-close pairs contribute to the analysis.

4. Matching Strategy Example

Example sequence for a single user:

timestamp	event	open_type
1000	open	manual
2000	open	auto
2000	close	

- The 'auto' open at timestamp 1000 overrides the earlier manual open (since no close occurred before it).
- The close at 3000 pairs with the auto open, creating one valid episode.

5. Findings

Open Type	Count	Mean Duration (s)	Median Duration (s)	Standard Deviation (s)
auto	1005	6,952.09	185.79	31,280.57
manual	625	1,672.80	12.28	11,817.02

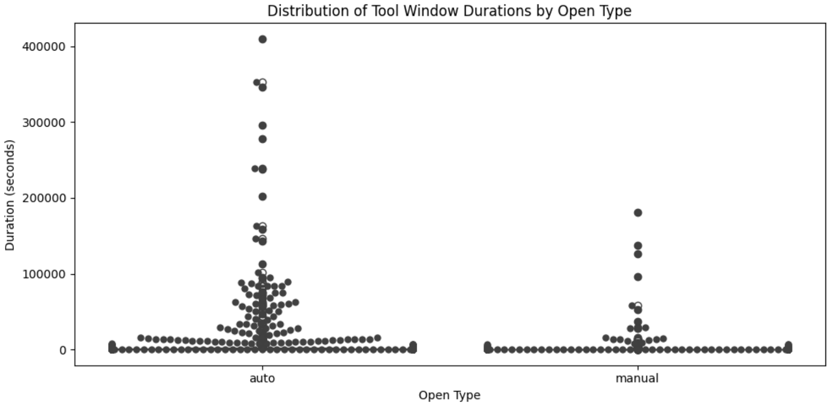
Statistical Test:

- Null Hypothesis (H_0): No significant difference between manual and auto durations.
- Alternative (H_1): A significant difference exists.
- Results: t-statistic = -4.825, p-value = 0.0000

As p-value < 0.05 → Null hypothesis (H_0) is rejected.

6. Visualizations

Distribution of Durations:



This visualization highlights duration patterns and support statistical findings.