Analysis Summary

In this assignment I analyzed a dataset containing information on the tool window activity, with information how the window was opened (manul or auto).

To work with the dataset I used the `pandas` library, the `matplotlib` library for data visualization, and `scipy.stats` for statistical analysis.

Because the raw dataset was not perfectly clean (for instance, a window closing event did not necessarily follow every opening event), it had to be pre-processed. My cleaning strategy was as follows:

1. First, I sorted all records by user identifier and, for each user, by timestamp. This allowed me to process the log of each user separately in chronological order.

2. In the second step, for each user I flagged all log events that did not have a clearly defined matching counterpart (the function `flag_unmatched_events_with_summary`). Such unmatched events may arise in three situations:

   (a) a window closing event without a preceding opening event;

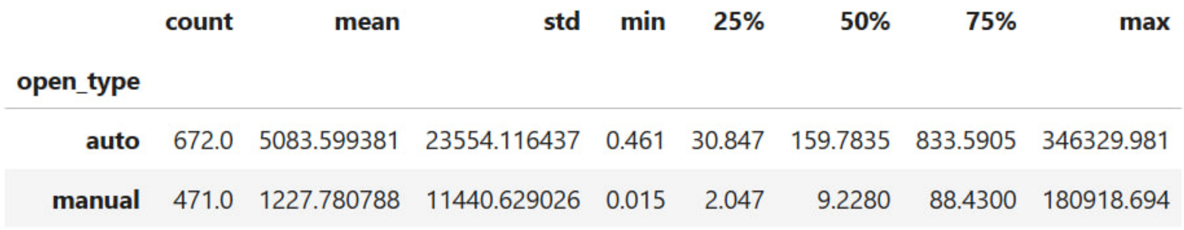
   (b) a window opening event without a subsequent closing event;

   (c) several opening events in a row. In this case it is impossible to determine which of the preceding openings each subsequent closing event corresponds to. Therefore, I labelled the entire sequence of openings and closings as unmatched until the number of closing events caught up with the number of openings.

3. Next, I removed all such "unmatched" events from the log and, for every remaining opening–closing pair, computed the window *lifetime* (session duration).

4. Finally, I split the resulting sessions into two groups depending on the type of window opening (see Annexe 1) and plotted the corresponding distributions (see Annexes 2 and 3). I then performed statistical tests to determine whether the difference in session durations between the two variants is statistically significant. I first applied the Shapiro–Wilk test to assess whether the distributions were approximately normal. Since normality was rejected, I used the Mann–Whitney U-test, which showed that the difference in session durations between the variants is statistically significant.

Annex 1. Table of Session Durations by Variant



	count	mean	std	min	25%	50%	75%	max
open_type								
auto	672.0	5083.599381	23554.116437	0.461	30.847	159.7835	833.5905	346329.981
manual	471.0	1227.780788	11440.629026	0.015	2.047	9.2280	88.4300	180918.694

Figure 1: Table of session durations by variant (manual vs. auto).

Annex 2. Distribution of Session Durations by Variant

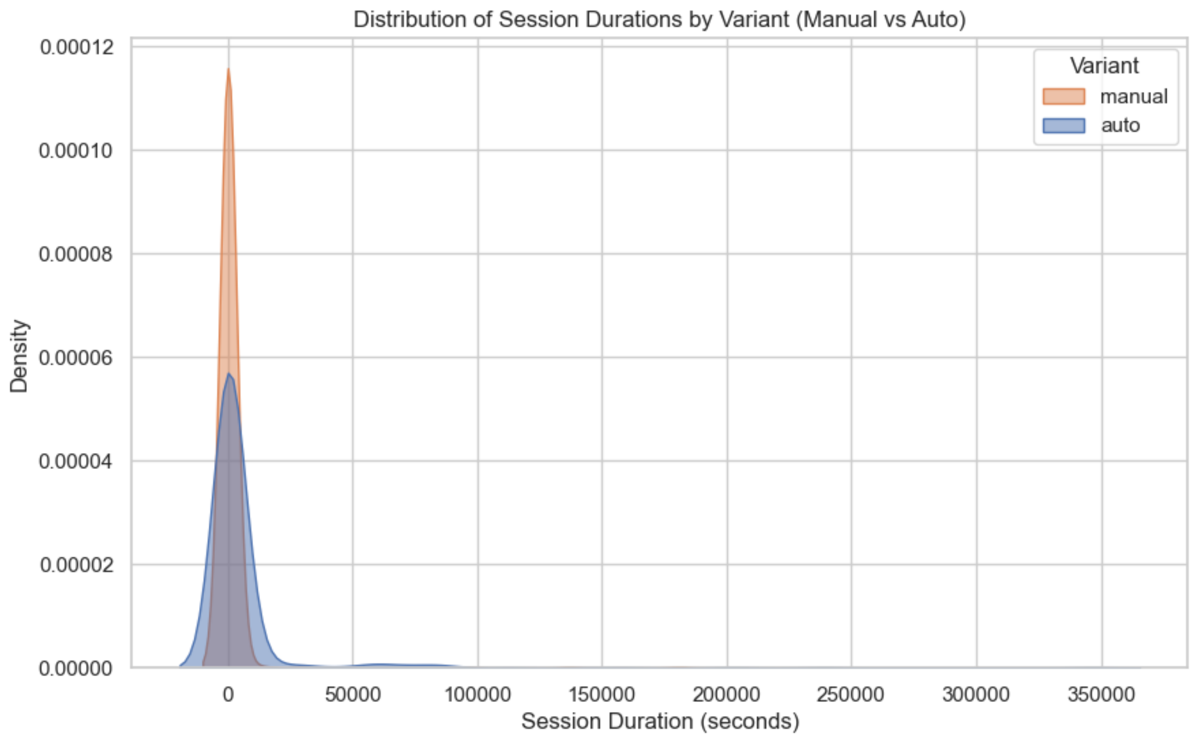


Figure 2: Distribution of session durations by variant (manual vs. auto).

Annex 3. Session Durations by Variant

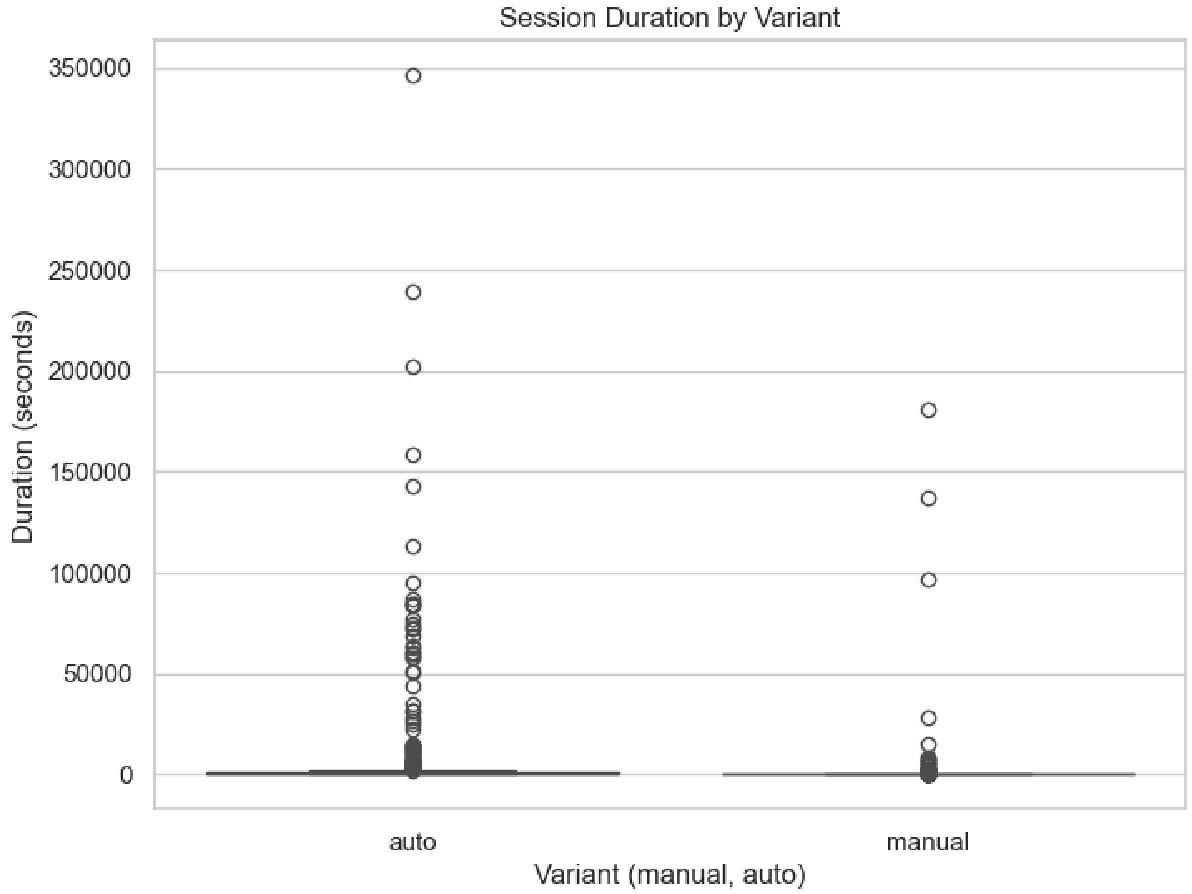


Figure 3: Session duration by variant (manual and auto).