

TYPING PERFORMANCE ANALYSIS

Submitted by:

Manav Maharishi (2024010057)
Riya (2024010089)

MCA 2nd Year

Submitted to:

Dr. Anjula Mehto
Assistant Professor



Computer Science and Engineering Department
Thapar Institute of Engineering and Technology, Patiala

November 2025

TABLE OF CONTENTS

S. No	Topic	Page No.
1	Introduction or Project Overview	3
2	Problem Statement	4
3	Overview of the Dataset used	5
4	Project workflow	6-7
5	Results	8-13
6	Conclusion	14
7	GitHub Link	15

Introduction or Project Overview

This project conducts a comprehensive analysis of long-term typing-test data to examine performance trends and behavioral patterns associated with words-per-minute (WPM). The methodology involves systematic preprocessing of timestamped sessions, transformation of the data into analytically meaningful structures, and the development of a visual analytics dashboard. Compressed-time scatter plots are employed to represent WPM and accuracy without distortion from irregular testing intervals, enabling clearer observation of progression over time. Additionally, a Monkeytype-style activity heatmap is constructed to capture day-level and week-level test frequency, revealing temporal habits and consistency of practice. Together, these visualizations highlight meaningful relationships between accuracy, speed, and usage patterns, while offering an interpretable and structured view of user behavior. The resulting dashboard provides a nuanced understanding of typing performance and supports data-driven insights into factors contributing to improvement.

Problem Statement

The primary objective of this project is to analyze long-term typing-test session data in order to identify patterns, trends, and behavioral factors that influence typing performance. The dataset consists of timestamped records containing words-per-minute (WPM), accuracy, raw speed, and test-level metadata. Due to irregular testing intervals and varying daily activity, a key challenge is to represent performance trends without distortion from inactive periods. This project addresses this by applying preprocessing steps to structure the temporal data and by developing a visual analytics dashboard that summarizes performance over time. Compressed-time scatter plots are used to illustrate WPM and accuracy trends, while a Monkeytype-style activity heatmap captures weekly and monthly patterns of test frequency. These visualizations facilitate the identification of improvements, plateaus, and fluctuations in typing behavior. The problem addressed by this project is to transform raw, irregular session logs into interpretable insights that reveal how consistency, practice habits, and accuracy relate to changes in typing speed over time.

Overview of the Dataset used

Dataset: A collection of 1000 typing-test session records containing timestamped performance metrics and behavioral indicators. The dataset documents real user typing behavior across multiple days and varying test durations, enabling temporal trend analysis, activity-pattern visualization, and performance interpretation.

Feature List

- **wpm:** Target variable; words per minute recorded for the test.
- **acc:** Test accuracy (percent or fraction); higher values tend to link with higher effective WPM.
- **rawWpm:** Raw typing rate before adjustments; shows uncorrected speed.
- **consistency:** Within-test stability measure; larger values indicate steadier performance.
- **testDuration:** Test length in seconds; affects measurement reliability and comparability.
- **afkDuration:** Seconds away from keyboard during the test; many zeros, better shown with counts or histograms.
- **incompleteTestSeconds:** Seconds recorded for incomplete or abandoned tests; concentrated at zero with a long tail.
- **restartCount:** Number of restarts during the test; integer and often zero, best visualized with count plots.

Project Workflow

1. Data Loading

The workflow begins by loading *results.csv* into a pandas DataFrame and performing initial validation steps such as inspecting the first few rows, checking the dataset shape, and verifying datatypes. Missing values and malformed timestamps are identified at this stage to ensure that subsequent processing is based on correctly structured data.

2. Feature selection

After loading, the relevant features are selected for analysis. These include WPM, accuracy, rawWpm, consistency, testDuration, afkDuration, incompleteTestSeconds, restartCount, and the derived time and date columns. Selecting these columns early ensures that the analysis stays focused on the session-level metrics needed for meaningful interpretation.

3. Cleaning & imputation

Data cleaning involves removing rows where timestamp conversion fails and ensuring that all selected numeric features can be safely used for visualization. Although the notebook primarily relies on dropping malformed rows rather than heavy imputation, the process ensures that every plotted value corresponds to a valid recorded session without fabricated or unreliable entries.

4. Exploratory data analysis (EDA)

We compute analysis and provides an initial understanding of performance patterns. Summary statistics are computed for key metrics such as WPM, accuracy, and rawWpm. Visual exploration is carried out using histograms, scatterplots, and comparisons like rawWpm versus effective WPM, which help reveal value ranges, variation, and stability within the dataset.

5. Handle skew and zero-inflation

Some behavioral variables, such as afkDuration, incompleteTestSeconds, and restartCount, show strong zero-inflation and irregular distributions. These characteristics are handled by choosing visualization techniques that accurately reflect their discrete nature, such as count-based plots and carefully scaled histograms, which prevents misleading interpretations that could arise from continuous-style summaries.

6. Time Handling for Trend plots

Because the dataset contains long gaps between typing sessions, using actual calendar dates on the X-axis leads to distorted plots with large empty spaces. To address this, the workflow adopts a compressed-time approach where each session is evenly spaced along the X-axis but still labeled with real dates.

7. WPM Trend Visualization

Using the compressed-time structure, the notebook generates a scatter plot of WPM across sessions, with cross-shaped markers that make each individual test easily identifiable. A faint line connects the points to help highlight overall progression. Selected date labels are placed along the X-axis to anchor the timeline without overcrowding the figure.

8. Accuracy Trend Visualization

A second compressed-time plot is generated for accuracy, preserving the same clean layout and sampling of date labels. This parallel visualization allows direct comparison between the evolution of WPM and accuracy while maintaining stylistic and structural consistency across plots.

9. Activity Heatmap Construction

To analyze practice frequency and typing habits, the workflow builds a Monkeytype-style activity heatmap. A complete daily date range is generated to avoid missing gaps, merged with actual session counts, and reshaped into a weekday-by-week matrix. The resulting heatmap visually highlights active periods, inactivity streaks, and overall patterns of engagement throughout the analysis window.

10. Diagnostics Observations and Limitations

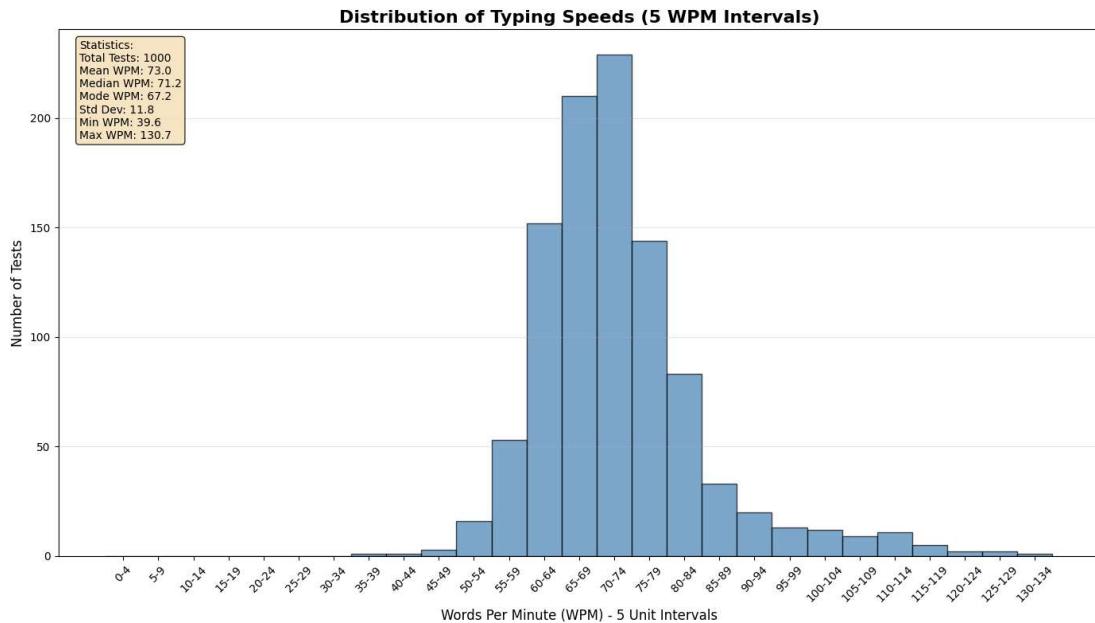
Throughout the workflow, the notebook documents reasons for the chosen visualization strategies—particularly the necessity of compressed time due to irregular activity. It also notes dataset limitations such as skewed distributions, zero-heavy fields, and the absence of continuous daily activity, ensuring the analysis remains transparent and academically grounded.

11. Final Outputs and Interpretations

The workflow concludes with a set of cleaned visualizations and summary files, including the WPM and accuracy trend graphs, the activity heatmap, and optional rolling-average views. These outputs collectively transform raw typing logs into interpretable visual insights that describe performance evolution, behavioral consistency, and long-term activity patterns.

Results

- **Performance distributions (Histograms):** Histograms for WPM, rawWpm, and accuracy show clear patterns in overall performance. WPM values cluster around a stable range with occasional peaks, while rawWpm is more skewed, indicating faster uncorrected speed during certain sessions. Accuracy remains consistently high for most tests, with few low-accuracy outliers. These distributions provide the baseline understanding needed to interpret long-term typing behavior.
- **Time-based Trends (Scatter plots):** Since the dataset includes irregular and widely spaced typing sessions, compressed-time plots were used to avoid misleading gaps. The WPM trend plot displays test-by-test progression using evenly spaced session indices but retains actual dates as selected x-axis labels. This reveals genuine improvements and performance dips without distortion from long inactivity periods. A similar plot for accuracy shows how correctness fluctuates alongside speed, helping identify sessions where performance shifted due to accuracy changes.
- **Raw vs Effective Speed Comparison:** A direct comparison between rawWpm and final WPM illustrates how accuracy adjustments affect observed performance. Sessions with high errors show larger gaps between raw and effective speed, while consistent sessions show nearly identical traces. This provides insight into how typing behavior and accuracy jointly determine final speed.
- **Activity Heatmap (Monkey Type style):** A calendar-style heatmap displays daily typing activity across weeks and weekdays. By constructing a full date grid and mapping session counts, the heatmap highlights streaks of continuous practice, sparse activity periods, and preferred practice days. Weekly blocks and month labels make the visualization easy to interpret, closely mirroring the visual style used by Monkeytype for activity tracking.
- **Behavioral Feature Visualizations:** Zero-inflated fields such as afkDuration, restartCount, and incompleteTestSeconds were visualized using count-based plots rather than continuous boxplots. These graphics clarify how often these behaviors occurred and prevent distortion caused by large numbers of zeros. They also support interpretations of typing behavior, such as tendencies to restart or abandon tests.
- **Daily Summary Insights:** A per-day aggregation (saved as *daily_summary.csv*) summarizes average WPM, accuracy, and number of sessions each day. This makes it possible to correlate practice intensity with performance outcomes and identify days with unusually high or low performance.



Typing Speed Distribution Analysis

The histogram of typing speeds shows a distinctly positively skewed (right-skewed) distribution. The most prominent feature of the graph is the large concentration of tests in the 60–79 WPM range, forming a high central block where the majority of attempts lie. This indicates that this interval represents your typical and most stable typing performance.

To the right of this central cluster, the graph displays a gradually thinning tail extending from 80 WPM all the way to the 130–139 WPM interval. The frequency of tests decreases steadily as speeds increase, which creates the long right tail that characterizes positive skewness. This shows that high-speed performances do happen, but they are significantly less frequent and occur only under ideal conditions such as familiarity with the prompt, higher focus levels, or lower fatigue.

On the left side of the distribution, the graph is almost flat — the 0–50 WPM intervals have very few attempts. This lack of a left tail reinforces that your lower-speed performances are rare, and that the skew comes entirely from occasional high-speed peaks, not slow outliers.

Mean–Median–Mode Structure in This Graph

Because the bulk of the bars are tightly grouped around 60–79 WPM, the mode of the distribution lies in this region.

However, the mean is pulled to the right due to the long tail of 90+, 100+, and 110+ WPM tests.

The median sits between the two, slightly higher than the mode but noticeably lower than the mean.

This mode < median < mean pattern is visually consistent with the shape seen in the graph.

How This Skew Changes Over Time

As practice sessions accumulate, two visible changes typically occur in graphs like this:

The central cluster slowly shifts right.

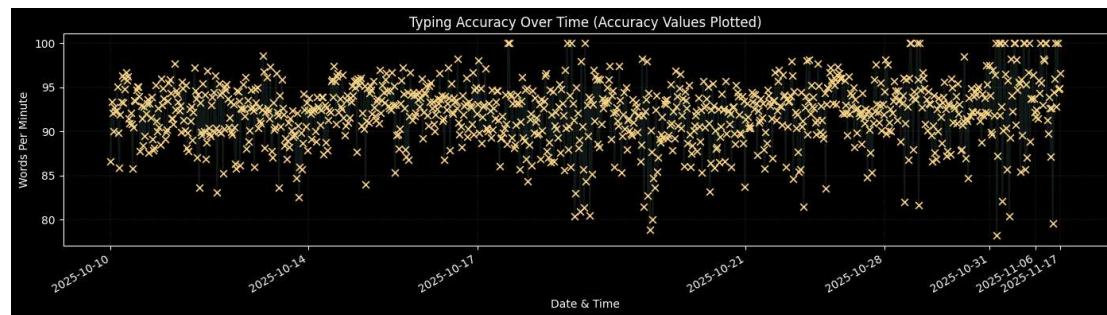
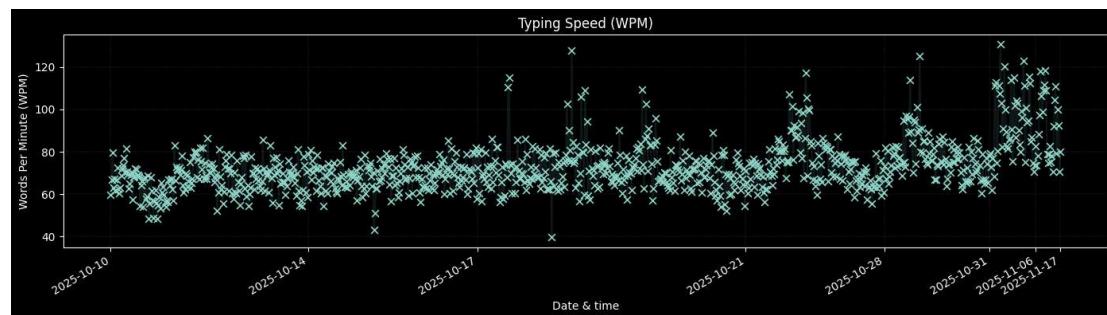
If your normal typing speed improves from 60–70 WPM to 70–80 WPM, this shift will be clearly reflected in a new peak forming in the next interval.

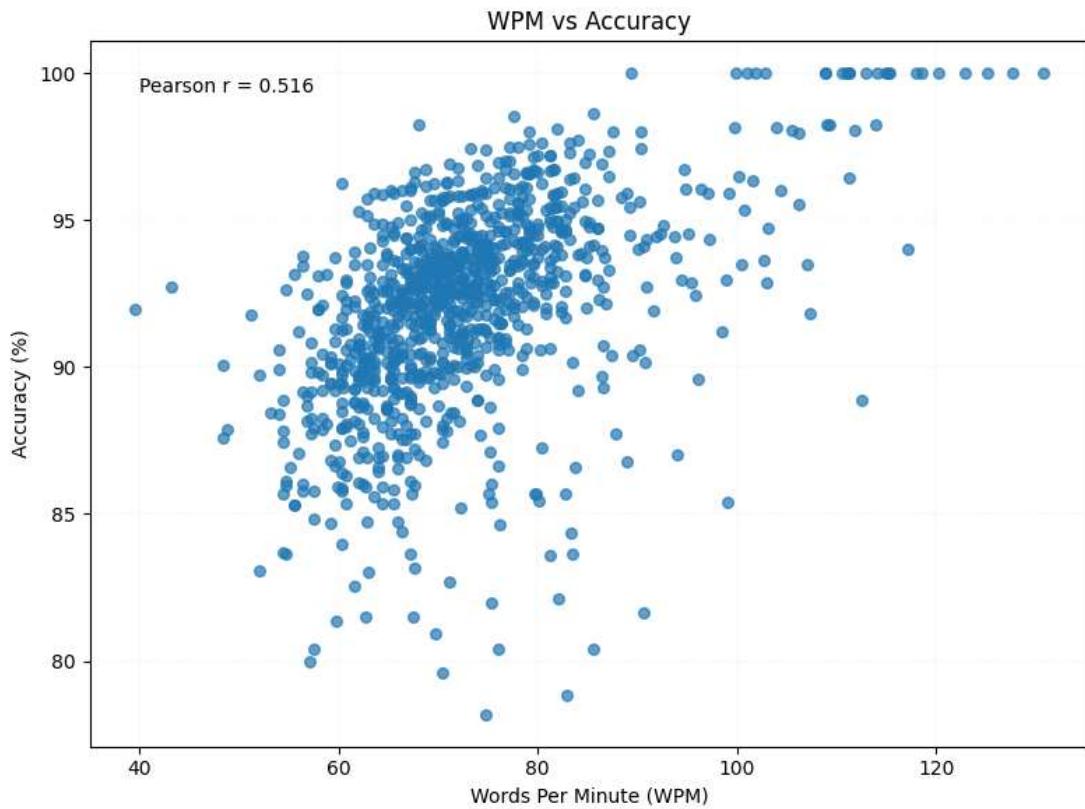
The right tail becomes less extreme.

As your higher-speed attempts become more frequent and less “rare,” the tail no longer stretches thinly across many intervals.

Instead, the distribution becomes tighter and more symmetrical, indicating stabilization and improvement in your baseline skill.

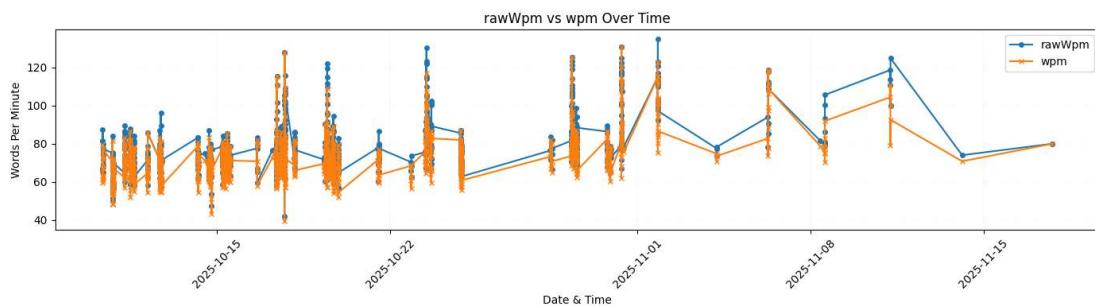
Together, these trends visually demonstrate not only growth in typing speed, but also consistency, as your performance transitions from a right-skewed distribution to a more balanced shape centered at a higher WPM.

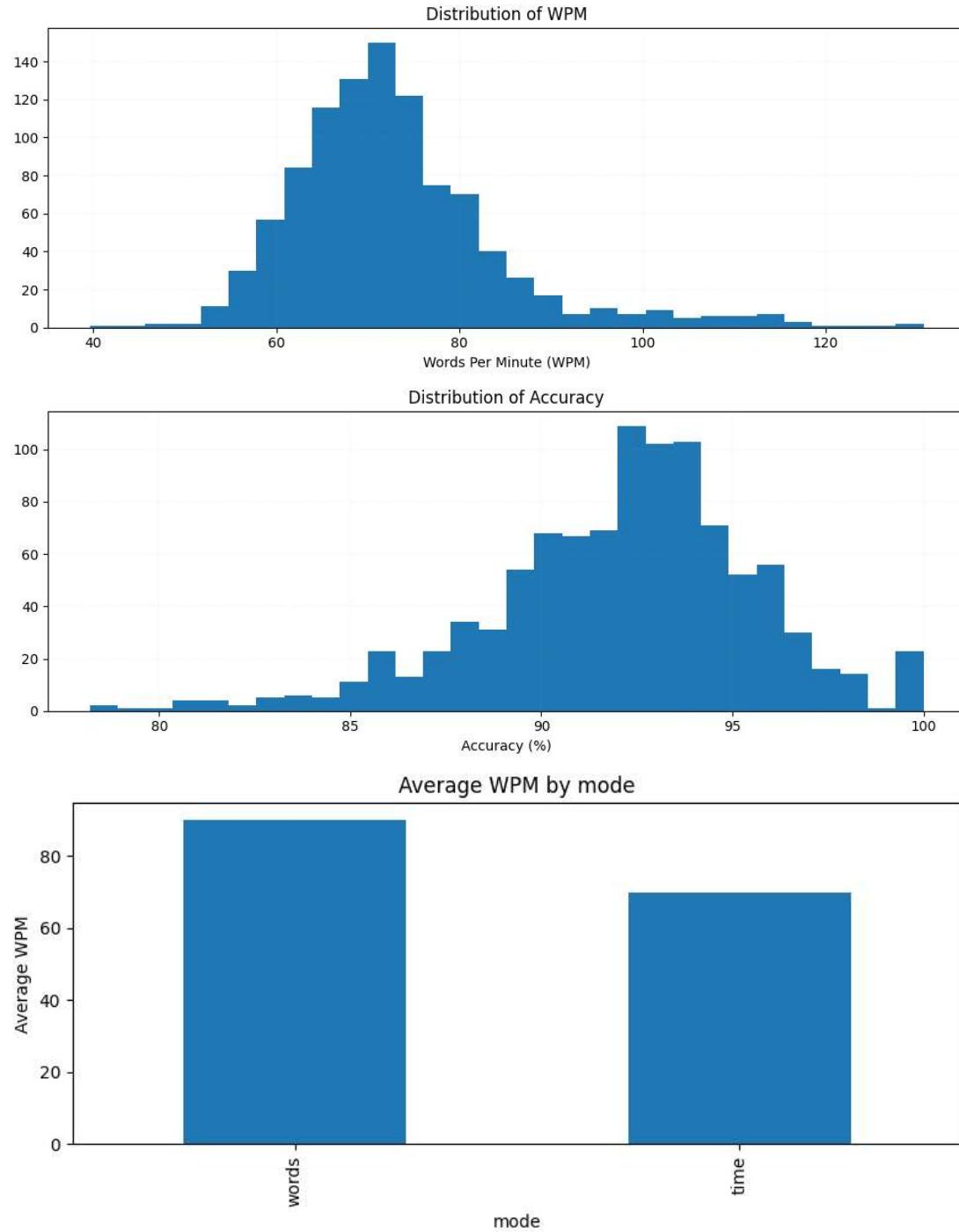




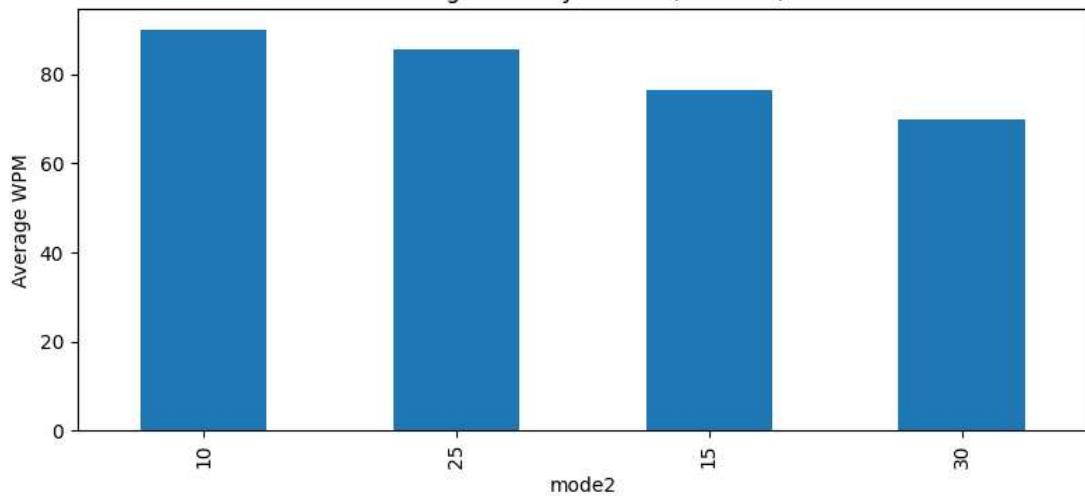
An r-value of 0.516 indicates a moderate positive correlation between WPM and accuracy.

In simple terms, as typing speed increases, accuracy also tends to increase — but the relationship is not extremely strong. It means faster typists generally maintain better accuracy, but there is still quite a bit of variability: some fast typists may be less accurate, and some slow typists may be very accurate. Overall, the trend is upward, showing that speed and accuracy improve together to a reasonable extent.





Average WPM by mode2 (duration)



Conclusion

The analysis of 1000 typing sessions provides a clear picture of how performance and behavior evolve over time. Compressed-time visualizations reveal gradual improvement in words-per-minute and highlight short-term fluctuations linked to changes in accuracy. The comparison between rawWpm and effective WPM shows how accuracy directly influences final speed, especially during inconsistent sessions. The activity heatmap further illustrates practice patterns, identifying active streaks, periods of inactivity, and preferred days of engagement. Together, these visualizations successfully convert raw session logs into interpretable insights, offering a well-rounded understanding of typing progress, consistency, and user behavior

GitHub Link

Access the dashboard and source code:

[MCA308_Assignments/typing_dashboard.ipynb at main · Riya-dudeja/MCA308_Assignments · GitHub](https://github.com/Riya-dudeja/MCA308_Assignments)