Attendance Analytics Starter

From raw logs to actionable insights

By Rashidat Sulyman

AI

DATA

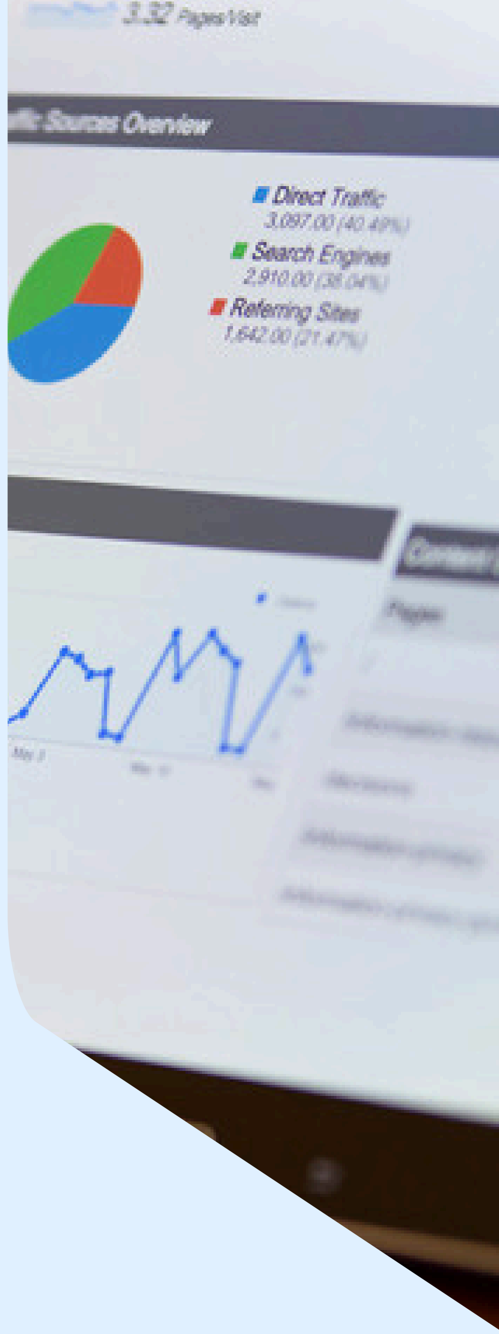
EDUCATION

Introduction

Virtual learning generates vast amounts of attendance data, but raw logs alone rarely provide meaningful insights for instructors or coordinators. This project transforms those logs into actionable analytics by introducing **grouped aggregation** and **simple predictive thinking**.

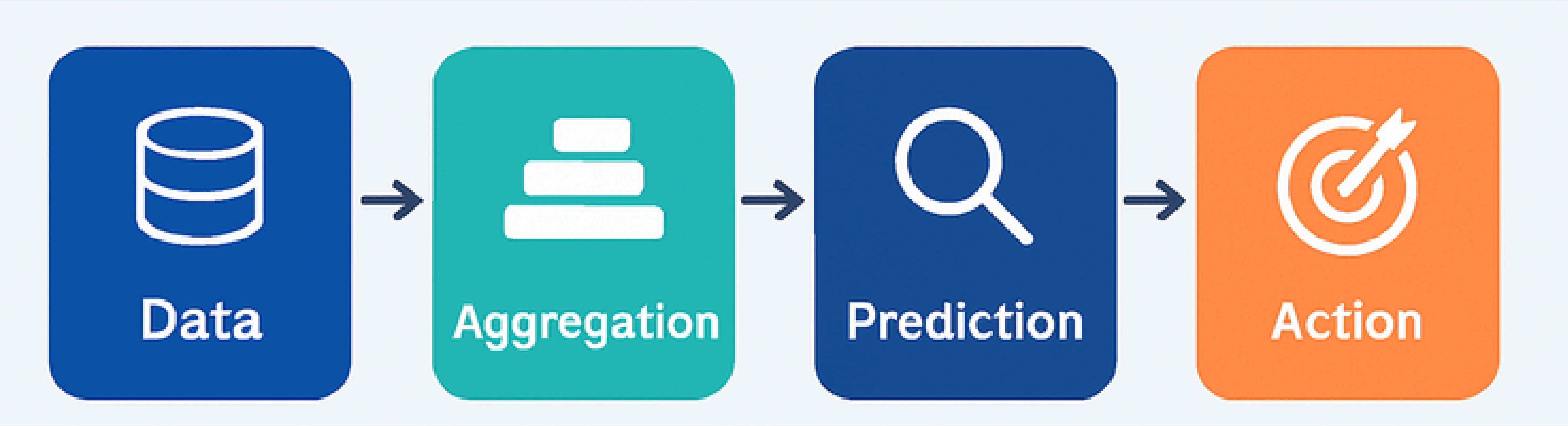
Through grouped aggregation, attendance records are summarized into clear measures such as sessions attended per student, average participation duration, and the most active days of the week. These summaries allow instructors to quickly identify patterns of engagement across individuals and cohorts.

Simple predictive thinking is applied by flagging students with low attendance rates or declining participation trends. Using rule-based thresholds, the project highlights learners at risk of disengagement, enabling coordinators to intervene early and improve outcomes.
The result is a starter framework for **Attendance Analytics**: a practical tool that bridges data science fundamentals with AI-inspired insights, empowering educators to move from raw data to proactive decision-making in virtual classrooms.



Objectives

The project aimed to transform raw virtual attendance logs into actionable insights, by analyzing student engagement patterns and session metrics. This will support instructors and coordinators in monitoring student engagement, to flag students with low engagement for early intervention and improve virtual learning.



Data Exploration >

Data Set

Source:

- Virtual attendance logs (session_id, student_id, joined_at, left_at) from online sessions from GoMask.ai data sets . It was downloaded into a desktop folder which was later loaded into pandas dataframe for analysis.

Key Variables:

- student_id: Unique identifier for each student
- session_id: Unique identifier for each session
- join_time and leave_time: Timestamps of attendance
- duration_minutes: Computed session duration

Data Set

Screen-shot of Pandas-profiling Reports

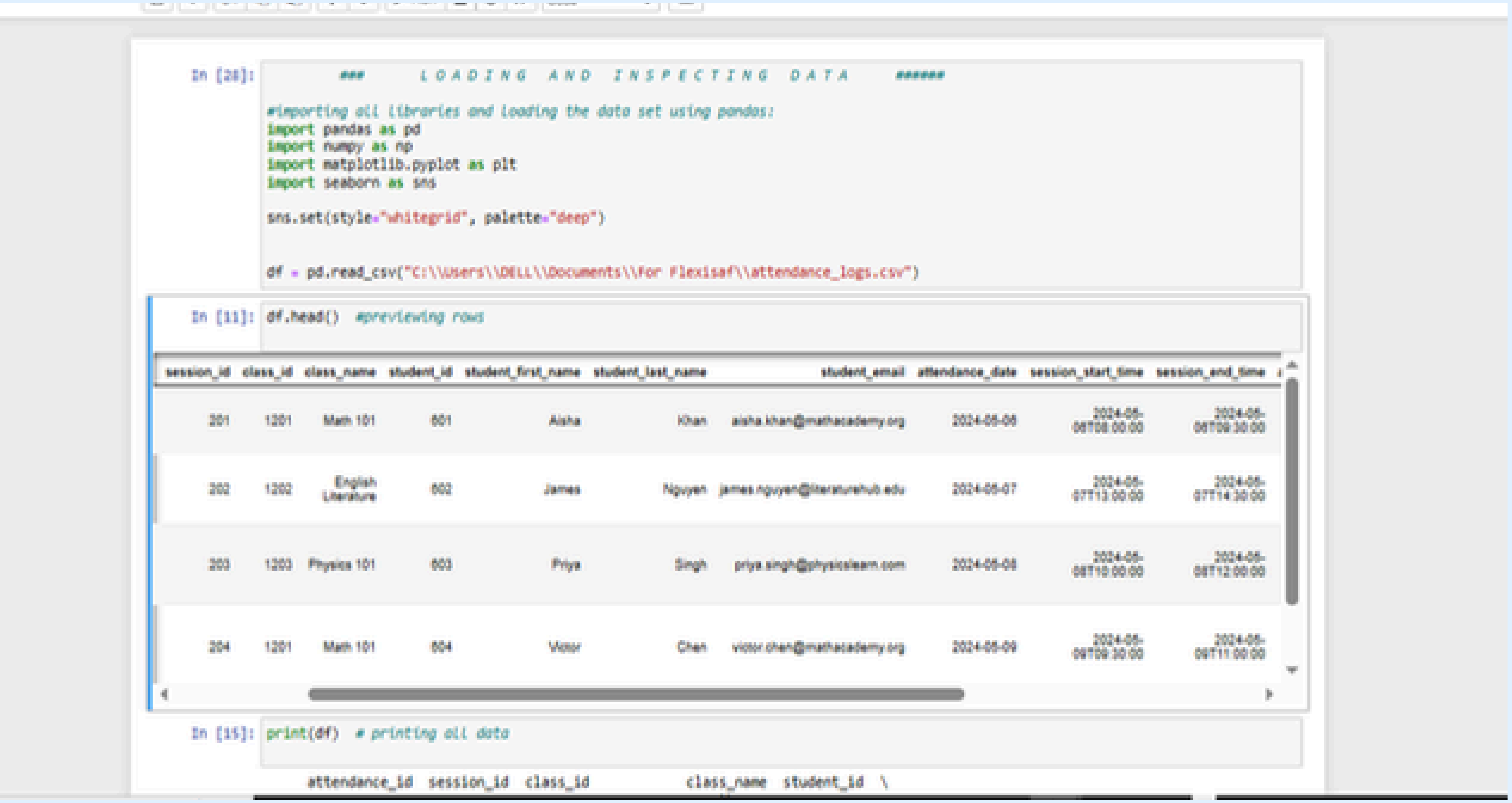


Fig 1: Loading the Dataset in Pandas Dataframe

Methods

Data Cleaning & Processing

- Removed missing values.
- Converted timestamps (join_time, leave_time) into proper datetime format.
- Computed duration of attendance(session durations) by subtracting join time from leave time.
- Derived additional fields such as date and day of week and hour features.

Feature Engineering

Derived Metrics:

- **sessions_attended:** Number of sessions attended per student
- **avg_duration_min:** Average session duration per student
- **attendance_rate_pct:** Percentage of sessions attended
- **low_engagement_flag:** Rule-based indicator of disengagement

AI Feature

Initial Engagement Rule

- **Threshold:** Students attending < 50% of sessions → flagged as low engagement.
- **Outcome:** 140 students flagged (100% of dataset).
- **Issue:** Too rigid. Since most students attended only 1 session, nearly everyone was flagged, making the rule uninformative.

Fixed Attendance & Duration Thresholds

- **Thresholds:** Attendance Rate < 60% → flag low attendance. Avg. Duration < 20 minutes → flag low duration.
- **Outcome:** Many students flagged for low attendance, but not for duration (most stayed ~90 minutes).
- **Issue:** Attendance rate calculation was skewed (values ~1.2%), so nearly all students were flagged low attendance. Duration threshold was too low to differentiate meaningfully.

Median-Based Dynamic Thresholds

- **Thresholds redefined as dataset medians:**Attendance Threshold = 1.2% (median attendance rate) Duration Threshold = 88.9 minutes (median session duration)
- **Outcome:** Flags balanced: ~49% low engagement, ~51% consistently engaged.
- Distribution became more realistic, separating students into meaningful categories.
- Strength: Adaptive to dataset characteristics and avoids blanket-flagging everyone.

Machine Learning Component

Model : Random Forest Classifier

- Features: sessions_attended, avg_duration_min, attendance_rate_pct
- Target: low_engagement_flag
- Evaluation: Train/test split and 5-fold cross-validation.

Result:

- Class 0 (Engaged students)
  - Precision = 1.00 → Every student predicted as engaged was truly engaged.
  - Recall = 1.00 → The model correctly identified all engaged students.
  - F1-score = 1.00 → Perfect balance between precision and recall.
  - Support = 22 → There were 22 engaged students in the test set.
- Class 1 (Low-engagement students)
  - Precision = 1.00 → Every student predicted as low-engagement was truly low-engagement.
  - Recall = 1.00 → The model correctly identified all low-engagement students.
  - F1-score = 1.00 → Perfect balance.
  - Support = 20 → There were 20 low-engagement students in the test set.
- Overall accuracy = 1.00 (100%)
  - Out of 42 students in the test set, the model classified every single one correctly.
- Macro avg & Weighted avg = 1.00
  - Both averages across classes are perfect, meaning the model treated both classes equally well.
  - Features: sessions_attended, avg_duration_min, attendance_rate_pct
  - Target: low_engagement_flag
  - Evaluation: Train/test split and 5-fold cross-validation.

Visualization

Used Python (Pandas, Matplotlib, Seaborn) to generate charts for trends and comparisons.

Key Findings

Everything you need to know in
just a few scrolls

Part 1

Attendance Patterns

Attendance Patterns

Most Active Days of the Week

- **Peak activity**: Wednesday (25 sessions), Saturday (23), and Tuesday (23).
- **Lowest activity:** Sunday (12 sessions).
- **Finding:** Midweek and Saturday are optimal for scheduling important sessions. Mondays and Sundays show weaker participation, likely due to start-of-week fatigue and end-of-week disengagement

Attendance Patterns

Attendance by Date

- Attendance spread across 109 days, with occasional spikes (e.g., November 15: 4 sessions).
- Finding: Spikes may correspond to special events or deadlines. Regular attendance is steady but not concentrated.

Attendance Patterns

Time-of-Day Trends

- **Highest activity:** 9 AM and 1 PM (19 sessions each).
- **Secondary peaks:** 8 AM (15) and 10 AM (13).
- **Decline:** After 5 PM, with minimal activity by 8–9 PM.
- **Finding:** Students prefer morning and early afternoon sessions. Evening sessions are less effective.

Attendance Patterns

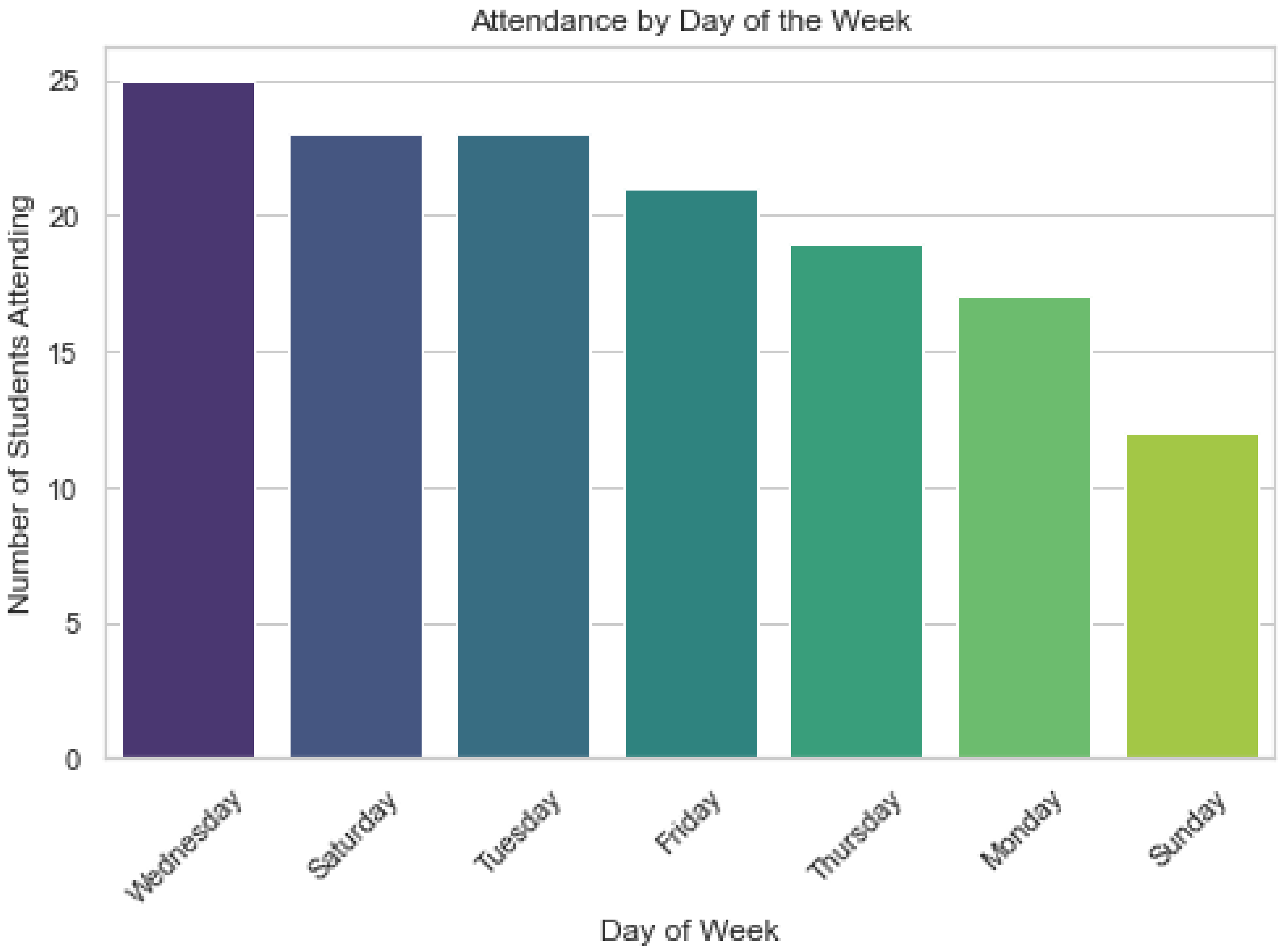


Chart 1: Most active days of the week

Attendance Patterns

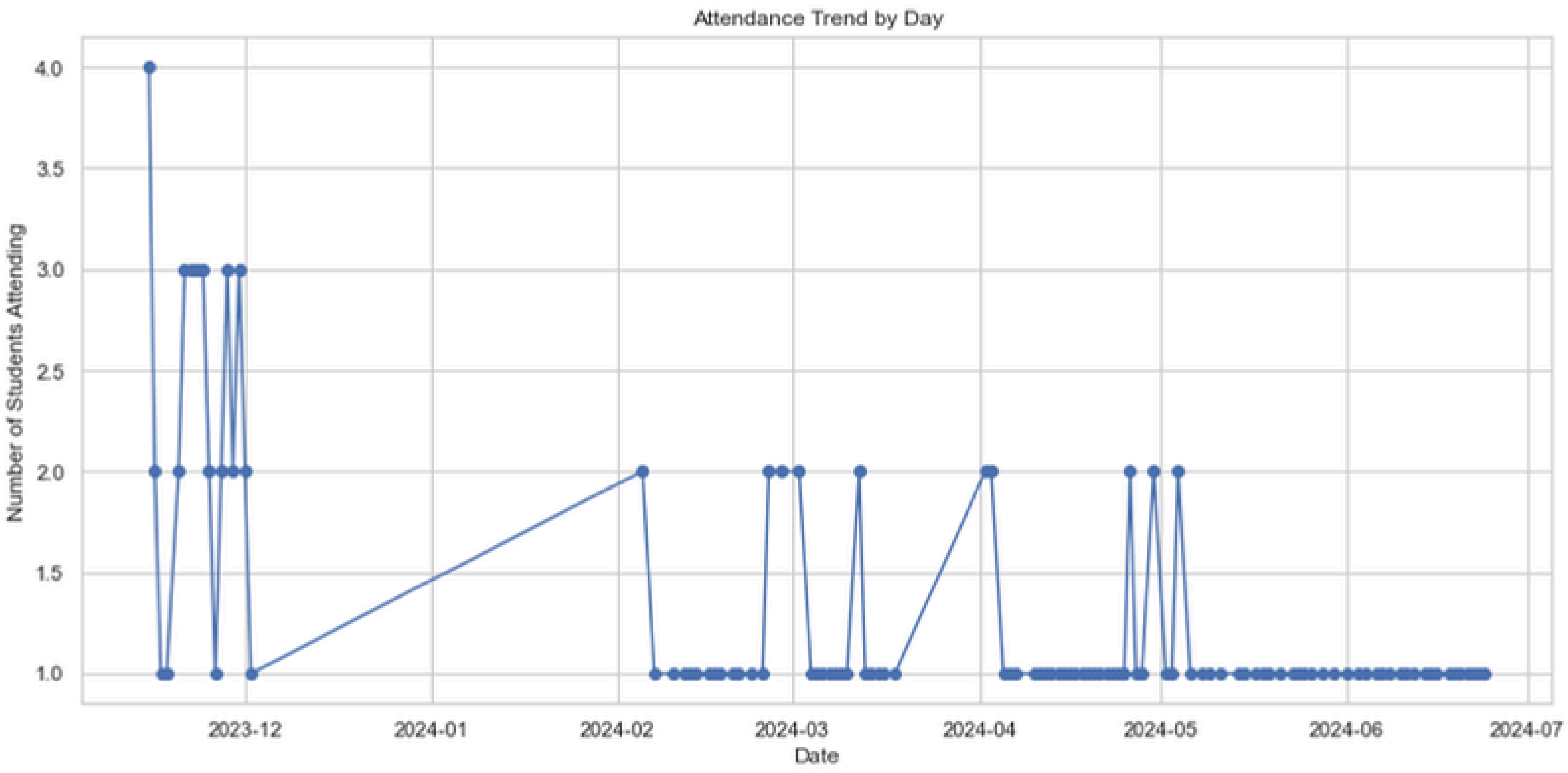


Chart 2: Attendace trend by date

Attendance Patterns

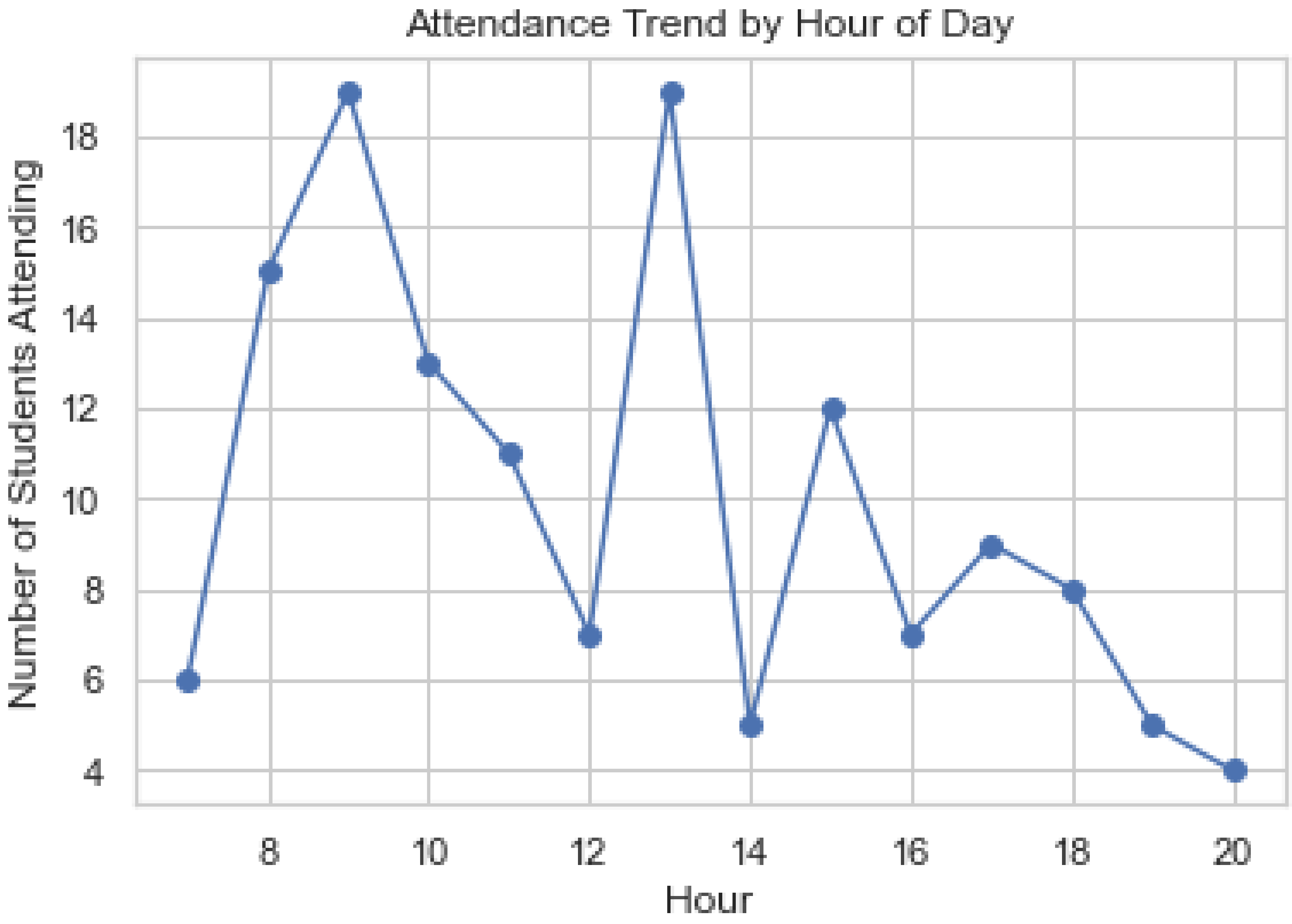


Chart 3: Attendance trend by hour

Part 2

Student Engagement Insights

Student Engagement Insights

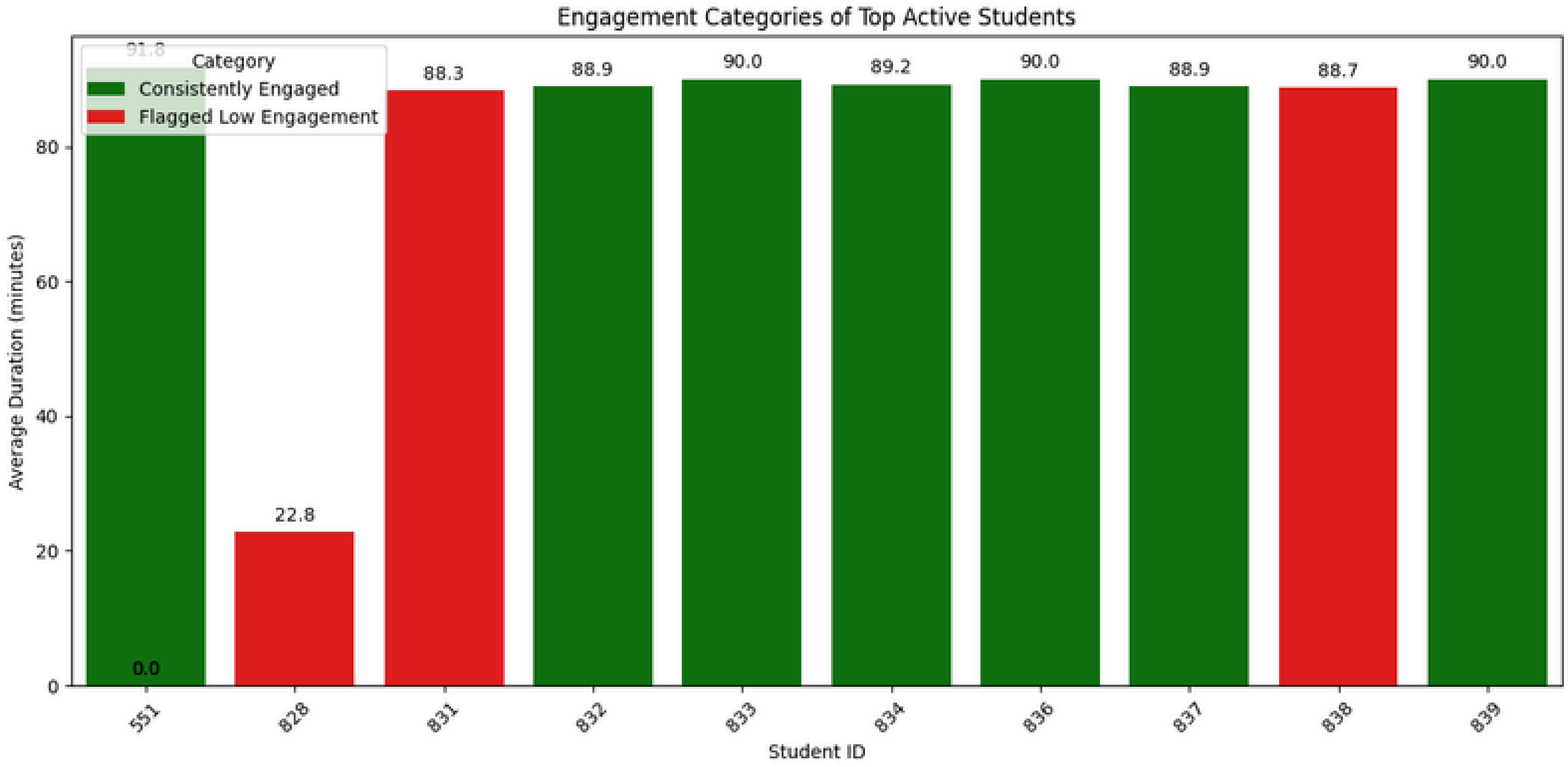


Chart 4: Bar chart of the top active students grouped by engagement category

Student Engagement Insights

- Most top active students are consistently engaged, staying for full sessions.
- A small subset attends but leaves early.
- Hence students were grouped into three categories:
  🟢 Consistently Engaged (~88–92 minutes average duration)
  🟠 Active but Short Duration (attend but stay < 60 minutes)
  🔴 Flagged Low Engagement (system-identified as disengaged)

Student Engagement Insights

Flagged Low engagement

- Successfully identified students with low average session durations or inconsistent attendance.
- Students 838, 831, and 828 were flagged as low_engagement_flag = True.
- This proves the system can distinguish between mere attendance and true engagement.

Student Engagement Insights

Consistently Engaged Students

- Students 551, 832, 833, 834, 836, 837, 839 show low_engagement_flag=False.
- Their average durations are ~88–92 minutes, which is healthy engagement and aligns with expected session length.
- These students represent the model group for consistent participation.

Student Engagement Insights

Active but short duration

- Some students attended sessions but stayed for shorter periods (< 60 minutes).
- Example: Student 828 attended but averaged only 22.8 minutes.
- These cases highlight the importance of measuring duration, not just attendance.

Part 3

Key Insights

Key Insights

Flagging works

The system correctly identified disengaged students.

Consistency matters

True engagement is defined by session duration, not just attendance.

Timing insights:

Midweek sessions drive the highest attendance; weekends show weaker participation.

Conclusion

The project demonstrates that engagement analysis must go beyond attendance counts. By combining metrics like session duration, attendance rate, and flagging logic, the system provides a clear, actionable picture of student behavior. This give educators a data-driven foundation to intervene early, reward consistent engagement, and optimize scheduling for better outcomes.

Recommendations

- Schedule critical sessions on Wednesday, Tuesday, or Saturday.
- Provide support or interventions for flagged students to improve engagement.
- Consider lighter or catch-up sessions on Sunday/Monday to balance attendance.