a) **Data Preparation**

The attendance data consisted of 27 handwritten JPG images, each representing a single class session with the class number, date, and a numbered list of student names and usernames. To improve extraction consistency, I performed lightweight but essential preprocessing. All images were resized to a standardized resolution to reduce noise and improve the OCR model's ability to identify handwritten text. I converted the images to grayscale, which helped enhance contrast between ink and background and resulted in more stable text extraction. The files were also renamed sequentially (e.g., class01.jpg, class02.jpg) to simplify batch processing and maintain a consistent mapping between image files and class sessions. No cropping or augmentation was required because key content was uniformly positioned across the sheets. This limited but targeted data preparation ensured more reliable extraction while preserving the original structure of the attendance forms.

b) **Model Creation and Extraction Pipeline**

To construct the attendance audit system, I used a pre-trained multimodal vision-language model (the Text Extractor GPT tool) as the foundation for automated extraction. This model was selected because it can interpret handwritten text embedded in images and output structured content when prompted appropriately. For each attendance sheet, I uploaded the image and used a strict JSON-format prompt specifying the fields to extract: class number, date, and a list of students with serial numbers, full names, and usernames. The standardized prompt served as a schema constraint that encouraged consistent formatting and minimized hallucination.

Because the model processes one image at a time, I used a hybrid workflow: automated extraction per image followed by automated aggregation in Python. After generating 27 JSON files, one for each class, I loaded them into a notebook and concatenated them into a single master dataset using Pandas. This dataset formed the basis for all subsequent analysis. To evaluate model correctness, I conducted manual verification on a randomly selected subset of classes, comparing the model-generated student counts against the handwritten sheets. Minor discrepancies were confined to unclear handwriting or missing usernames, but overall accuracy was high and sufficient for class-level statistical analysis. This approach qualifies as a pre-trained model pipeline supplemented with human-in-the-loop quality control.

c) **Results From Attendance Analyses Using the Model Outputs**

Everything can be found in the notebook titled ‘analysis.ipynb’ in the code directory, however, the results are also shown below:

(c.a) Number of classes held and their dates [10 points]

Across the dataset, 27 total class sessions were recorded. Each class number uniquely corresponded to a specific date, and the extracted class–date pairs were:

class	date
1	19 Aug 2025
2	21 Aug 2025
3	26 Aug 2025
4	28 Aug 2025
5	2 Sep 2025
6	4 Sep 2025
7	9 Sep 2025
8	11 Sep 2025
9	16 Sep 2025
10	18 Sep 2025
11	28 Sep 2025
12	25 Sep 2025
13	30 Sep 2025
14	2 Oct 2025
15	7 Oct 2025
16	14 Oct 2025
17	16 Oct 2025
18	21 Oct 2025
19	23 Oct 2025
20	28 Oct 2025
21	2025-10-30
22	Nov 4, 2025
24	2025-11-11
25	13 Nov 2025
26	18 Nov 2025
27	20 Nov 2025)

This confirms that attendance was successfully captured for all scheduled sessions.

(c.b) Median class attendance per class

Using the aggregated attendance counts from the extracted dataset, the median class attendance across all sessions was:

33 students per class.

This indicates that the typical class session had attendance at or near this value, with roughly half of the sessions above and half below this threshold.

(c.c) Dates with lowest and highest attendance [10 points]

The lowest attendance observed in the dataset was 14 students, which occurred on:

November 20, 2025.

The highest attendance recorded was 49 students, on:

August 21, 2025.

(c.d) Correlation between attendance and course evaluation dates; When attendance was highest [10 points]

To determine whether class attendance increased on course evaluation days, I labeled evaluation sessions as Class 15, Class 24, and Class 26, days where quizzes were held and grad paper presentations were made. The mean attendance for evaluation classes was:

40 students.

while the mean attendance for non-evaluation classes was:

32.7 students.

The correlation coefficient between the “evaluation class” indicator and attendance was:

r = 0.28.

Interpreting this value, there is a weak relationship between evaluation timing and higher attendance, but this could also be due to a lack of course evaluation dates (only 3 to look at). Intuitively, however, attendance is typically higher on these days.

The overall highest attendance across the entire dataset occurred on:

August 21, 2025,

showing that the highest turnout aligned more with earlier semester sessions rather than course evaluation dates. Perhaps students were more motivated to attend class earlier in the semester.

d) **Improvements With Additional Time**

With an additional week, several enhancements could significantly improve the performance and robustness of the attendance audit system. First, I could fine-tune a handwriting-aware OCR model (e.g., a lightweight Transformer or a Tesseract-based system retrained on our class's handwriting styles) to reduce extraction errors caused by ambiguous text. Second, I could implement a confidence-scoring mechanism that flags low-certainty fields for human review, creating a semi-automated verification loop. Third, batch processing could be fully automated by building a script that ingests all images at once and applies the text extraction model through an API instead of manually uploading each file. Finally, visual analytics and trend plots could be added to provide dashboards showing attendance patterns over time, variability, and correlations with key academic events. These improvements would enhance accuracy, scalability, and the overall value of the system.