Question 2 answers

a) To prepare the data, I first determined how to extract the text from the images. Initially, since the data could be seen as sensitive, I hoped to perform image extraction locally. To do so, I used EasyOCR, which was free to download. This was simple to run and returned a list of the text extracted, but I did have to shrink the images beforehand. However, the results were very poor, with names separated (that is, the first and last names would be listed independently, possibly in non-adjacent positions in the list) and with many misspellings and mixups that made it difficult to determine how many students there actually were. Given more time, I would have tried other programs I could run locally, such as Tesseract and PaddleOCR; if the data was very sensitive for an important project, I also could have tried paid programs like ABBYY FineReader.

   Instead, for the sake of time, I chose to use OpenAI's Text Extractor GPT, accessed through my university account. When using this, I also discovered that though the given files all had .jpg extensions, a majority were actually in MPO format, which the GPT could not read. As such, I used an online converter to convert the files from MPO to .jpg format (again, it would have been better to find a program to do this locally for sensitive data, but in the interest of time, I simply used a random website). I then was able to upload the files and get the text extracted. Interestingly, though I was able to upload the images as a zip file to the GPT, and it was able to extract text from them, the results were basically garbage. However, when I individually uploaded each image, the results were much better, with no obvious errors or problems. I therefore uploaded each image and asked for the results in a .csv file. I did notice that two of the images were the same (for class 25 on November 13th) and only uploaded one of the images to the GPT rather than duplicating the results (which I consider part of cleaning the data).

   In the interest of time, I chose not to perform any analysis involving tracking specific students; as such, I then manually compiled an excel file containing relevant statistics for each class. As I did this, I checked the AI's work by comparing its stated class number and dates to the contents of the relevant image, as well as confirming that the number of students was the same on both documents. I had to correct one class number/date, where the AI was clearly referencing the previous class numbers in its answers and therefore got confused by missing data (no data was provided for class 23, so the AI claimed that the data for class 24 was actually for class 23, despite the image clearly stating it was class 24). The number of students was always correct, and the AI automatically renumbered the students when the numbering was inaccurate on the original sheet (this may have been frustrating if the numbering meant something important, since the AI did not notify me that it did this). In my sheet, I also noted whether that class included an assessment (using the dates given in the final's description), whether the previous class was

successfully recorded, whether it was raining, whether the change to standard time had occurred yet, whether there was a holiday directly after the class, and how many classes remained until the next assessment. These are all factors I theorize may have impacted attendance; I collected all factors from course information except for the weather, which I found at weatherspark.com, and the holiday information, which I found using the USC official academic calendar.

b) To analyze the extracted data, I wrote code to manually perform the desired analyses locally. This could have theoretically been done by uploading my data to ChatGPT and asking it for the answers; however, due to the sensitivity of the data, I preferred local solutions. Additionally, since the requested answers are easily found using standard coding (or even could be done by hand if necessary), the introduction of AI into the process unnecessarily creates complexity and opportunities for unforced errors. As such, I chose to use a manual model coded and run locally rather than an AI model (either a locally-run or cloud-based one).

c)

i) There were 27 classes that occurred during the time covered by the dataset, though only 26 of those classes were included in the dataset (specifically, class 23 on November 6th is missing). The dates are all in the file AI_final_data.xlsx, and there are too many to easily list here, but they range from August 19th to November 20th.

ii) The median attendance per class is 33 students.

iii) The date with the lowest attendance is November 20th, with only 14 students (notably, this is both the class before Thanksgiving break, so some students may have left campus early, and the class on the due date of project 2, which I suspect some students may have been scrambling to complete and therefore skipped class). The date with the highest attendance was August 21st, with 49 students. This was only the second class, so many students likely had not yet decided to rely solely on the slides and recordings. However, I am unsure why more students came on the second day than the first - perhaps some students added the class right after the start of the semester?

iv) There is a correlation of higher attendance and assessment days: the mean attendance for assessments is 40 students per class, while the mean for the other classes rounds to 33 students per class. Additionally, the mean for assessments is dragged down by the attendance for the grad student presentations, which was only 34; presumably some undergraduates did not view this as an assessment for which they needed to show up. However, on the days when the entire class was being assessed (for quizzes 2 and 3), the mean attendance was 43. Only the first three classes of the semester had higher attendance than this, likely due to

students coming to learn about course policies and to determine whether they liked the live lectures more than just the recordings and slides.

d) To summarize my comments in part (a) about data preparation, if I had more time (and possibly a small budget), I would have attempted to run text extraction entirely on my device, rather than uploading the images first to a format converter and then to ChatGPT. This is because of the sensitivity of the data. Overall, however, I am very happy with ChatGPT's performance in extracting the text and would only prefer alternatives due to security concerns.

   I also would have tried to anonymize students and then analyze data at the individual level; in particular, I would have liked to determine a group of students who rarely came to class (perhaps at or below the 20th percentile for attendance) and a group of students who almost always attended (perhaps at or above the 80th percentile) and then compared the groups. When were the rarely-attending students more likely to attend class (probably the assessment days, but there could be other patterns)? When were the mostly-attending students more likely to be absent?

   Additionally, though as a fellow student I could never have this information, it would be interesting to compare the two groups, their grade averages, and perhaps their rates of engagement with the class by, for example, filling out the pre-course and mid-course surveys. I would also be interested to see if, among those who did fill out the pre-course surveys, there were any differences between answers (as a theory, perhaps pre-existing AI skill level affected attendance?) from the two attendance groups.

   Finally, given more time, I would have liked to perform an unsupervised clustering (perhaps K-means) algorithm (again run locally for data security) to examine the impact of other factors on class attendance. In my final data spreadsheet, I included information for a variety of other factors that I believe could impact attendance and therefore may be relevant to a clustering algorithm. However, best performance would require data from additional classes or semesters, because some of the factors (such as whether it was raining or whether the next day was a holiday) were very rare for the existing data, which means the analysis would likely struggle to find patterns. In this case, though, normalization would be needed to analyze across different classes and sections - perhaps this could be started by tracking attendance as a percentage of the total class rather than the absolute number.