Help Needed to Adjust Scanned PDF Documents

We're working on a project to digitize some historical records. We've had the records scanned into PDF documents, but in order to extract the data from the scanned PDFs, we need someone to help us adjust the documents a bit.

Here are more details about the project:

- We have about 30 PDFs that are scanned school rosters. The PDFs are in 7 different formats and the formats are identical within each PDF document.
- Since the pages are scanned, they are often at slightly different tilts.
- Here you'll find an example of each format. The resolution of the actual scanned files is higher than the blurred examples.
- The ideal output of your code will give us a PDF file for each of the input PDF files where all of the pages within the PDF are straightened (and perhaps cropped) so that they align to the same coordinates. This is so that we can draw a template of a grid in an OCR software package and apply the same grid to every page in the PDF.
    - In order for this template to work, the content must be fully contained in the cells of the grid so that the information in the top-right corner of the table, for example, is always inside the top-right corner cell of the grid.
    - Right now, as a result of scanning, each page is slightly different from the others, and if we use a grid template, we lose some information because the grid lines are on top of text or information ends up in the wrong cell.
    - Some formats have more whitespace between columns than others, but this is consistent across pages within the PDF once the pages are straightened.
    - Each format has a header that contains the labels for the data in each row/column. That header is always followed with at least one line. One potential way to straighten the pages is to identify that line and crop/straighten the page to it. However, previous attempts for this project have incorrectly cropped in the middle of the page, so data was lost. Additionally, one of the formats has lines between each row of data, so it's important to be careful that the code correctly identifies the header line and not the lines throughout the table. It's possible that you may have ideas to solve this issue that do not include using the header line, so we're absolutely open to alternate approaches.
    - Simply using a package to "deskew" the image has not solved the issue in the past.
    - In the past, we've attempted to use Houghlines for this project, and we ran into issues where the lines were incorrectly identified and we lost data from the tables.
- It's not necessary for your approach to include any OCR or text-scraping. We'll take the resulting PDFs and put them into our own OCR software.
- It's not necessary to draw a grid on each page, but it might be helpful for you to do so in your specific approach. Drawing a grid on each page may also eliminate the need to straighten/crop each page, however we'll need to make sure that our OCR software *always* recognizes the grid you draw on each page and that no information is lost when we run it through the OCR software.

By 12pm ET on Friday, August 2nd, please email scanned.pdfs.project@gmail.com with the subject line "Approach- [your name]" a summary of the approach you'd take to solve this issue.

Submissions submitted without this subject line or via methods other than the listed email address will not be accepted or reviewed.

You do not need to include lines of code in your approach, but rather tell us the steps your code would take and what transformations it would make to the pages in each document to align them all. Please also include possible errors you'd check for in your code and how you'd ensure that no data is lost when images are edited within your code *and* when the output PDFs are put in the OCR software. It's possible that you'll need different code for some/all of the formats, so please include details on the generalizability of your approach(es) to all/some of the formats.

We do not have a preference for coding languages/packages used, but you're welcome to reference the ones you'd use in your approach.

Based on the quality of approaches, we'll select a few candidates to write a sample program for two different formats. You'll be compensated $50 if you complete the sample program task *and* your resulting PDFs work when we put them into the OCR software. There will be additional work and compensation for those who successfully complete the sample task. **The sample program will need to be completed by 8/9/19.**