OCR image form to entities

Goal:
The goal is to run OCR on the images, get the text from the form, clean the text and perform simple NER to identify the key fields that we are looking for. The goal of the assignment is to assess the level of understanding and approach when solving a given problem.

Dataset:
Form 1:
https://drive.google.com/file/d/1ToSiLCI5zyOOLSoAY3Xpx7056V7ueh0r/view?usp=sharing
Form 2:
https://drive.google.com/file/d/1yoE3ToLTqojg10CmB1nquCO_-1VatFbO/view?usp=sharing

Steps & Expectations for the Candidate:

1. **Image Preprocessing**
    a. Load the given image forms.
    b. Apply necessary preprocessing techniques to make them suitable for OCR (e.g., thresholding, noise reduction).
2. **OCR Application**
    a. Extract text data from the processed images using an OCR solution of the candidate's choice.
3. **Text Data Cleaning**
    a. Process the extracted text to remove any artefacts or errors introduced during the OCR process.
    b. Format the cleaned text for further processing.
4. **Named Entity Recognition (NER) or key information extraction + Mapping**
    a. Use either a pre-trained model or design a solution to identify specific entities from the cleaned text, namely:
        i. Name
        ii. Travelling Date
        iii. Flight Number
        iv. Controlled Items Indicator (e.g., Yes/No, Checked/Unchecked)
    b. Candidates can leverage large language models or other ML techniques to build or improve their NER system.
5. **Data Storage**
    a. Design a schema or model to represent the extracted information.
    b. Store the processed data in a structured format, preferably a database or similar storage system. (eg: MySQL, SQLite or CSV)
6. **UI/UX**
    a. Implement a basic user interface where an individual can upload an image, see the extracted data, and manually edit it if needed. (eg: Streamlit, Gradio, or similar)

You will be required to work on a Jupyter notebook that uses libraries compatible with Python v3.8. You might want to set up a virtual environment for this. Do share with us the ipynb notebook file. Please make sure to comment your codes (section by section) on why you

chose this approach and explain it in simple words. Feel free to leverage on any 3rd party services or open source libraries.