2. PDF File Extraction

Task: Extract text from PDF and handle edge cases.

Explanation:
The goal is to retrieve text from a PDF while addressing challenges like:

- Scanned images instead of text.
- Multi-column layouts.
- Encoding issues.
- Handling empty or malformed pages.
- **Tools**: Libraries like PyPDF2, pdfplumber, or OCR tools like Tesseract for scanned PDFs.
- **Sample Code Snippet**:

python

Copy code

```
import PyPDF2
from pdfplumber import open as pdf_open


def extract_text(file_path):
    with pdf_open(file_path) as pdf:
        text = ""
        for page in pdf.pages:
            text += page.extract_text()
    return text
```