# Extracting Data from PDF to CSV format using Python

Michael Anderson, PhD

Director, Design and Computational Resources Unit Biostatistics and Epidemiological Research and Design (BERD) Oklahoma Shared Clinical and Translational Resources (OSCTR)

### PDF data

 Occasionally, data will be received in a pdf format like the one illustrated below making it difficult to use in standard statistical software packages.

Tools	Medicaid variables? X							R (	2	:
TOOIS	Wedicald variablesz ~								9 🐥	
P 🖶	• Q 🕒 Ć ⊄	× E 🖻	0 🗰 🕥 🕓	1 / 1 💦	💮 🕀 🕀	180% 🔹 🔛		, b	O_	2
_				-					_	
	Client	Client	Client Age	Race	Ethnic	Sex	First Date	Provider		
	(Deidentified)	Age at	(Completed	Group	Code	Code	of Service	(Deidentified)		
	. ,	DOS	Months)							
		(current	,							
		DOB)								
F	C#	0-18	0-250	Race	Hispanic	MorE	xx/xx/xxxx	D#	-	
	C <del>n</del>	0-18	0-230	Race			^^/ ^^/ ^^/	Г <i>#</i>		
					or Non-					
H					Hispanic				_	
	B#	19-25	251-470	Race	Hispanic	M or F	Yy/yy/yyyy	D#		
					or Non-					
					Hispanic					

This file is named "**Medicaid variables2.pdf**" The first row contains the column headings Column headings contain spaces and "(" Rows 2 and 3 contain the data

Like the column headings, the data contain special characters such as spaces, "-", and "/".

### Strategy: Read data into Python, output to

- Python will allow us to read in the .pdf and convert to .csv
- The special characters in the column headings and data of this dataset will need to be addressed as well. This may not be needed for your dataset.
- Python will allow us to easily find and change the special characters.
- The actual Python code required to do the above is minimal. However, for a first time use, there is some initial set up (downloading and installing software) required.
- Let's look at an overview of this process and then dive in to each step.

### Overview

#### **First Time Use**

- 1. Download and Install Java (requires admin approval)
- 2. Create a new JAVA\_HOME environment variable (**requires admin approval**)
- 3. Download and Install Anaconda 3 and Launch Spyder IDE
- 4. In the console type 'pip install tabula-py' and press enter

### Subsequent Use

- 1. Run python commands in Spyder
- 2. Check the exported .csv and adjust Python commands as needed.

Let's look at each of these steps in detail

### Download and Install Java (admin approval needed)

 Navigate to <u>https://www.java.com/en/download/</u> and select "Download Java"

jre-8u391-windows-x64.exe

64.4 MB • Done



By downloading Java you acknowledge that you have read and accepted the terms of the Oracle Technology Network License Agreement for Oracle Java SE

2. When the file install

is downloaded, click to open and

Note: If you don't have administrator privileges for your computer, you will need the admin's permission (they'll need to login) to do this.

1. Launch "Control Panel" from Windows System folder



### 2. Select System and Security



### 3. Select System



### 4. Select Advanced Settings (right side of settings page)



### 5. Select "Environment Variables" button

System Properties

Computer Name Hardware Advanced System Protection Remote										
You must be logged on as an Administrator to make most of these changes.										
Performance										
Visual effects, processor scheduling, memory usage, and virtual memory										
Settings										
User Profiles										
Desktop settings related to your sign-in										
Settings										
Startup and Recovery										
System startup, system failure, and debugging information										
Settings										
Environment Variables										
OK Cancel Apply										

 $\times$ 

### 6. Select "New" from the System variables box

Variable	Value								
OneDrive	C:\Users\mander10\OneDrive								
Path	C:\Users\mander10\AppData\Local\Programs\Python\Python								
TEMP C:\Users\mander10\AppData\Local\Temp									
ТМР	C:\Users\mander10\AppData\Local\Temp								
	New Edit Delete								
/stem variables	Value								
ComSpec	C:\WINDOWS\system32\cmd.exe								
DriverData	C:\Windows\System32\Drivers\DriverData								
MINITAB_LICENSE_FILE	C:\ProgramData\Minitab\Minitab 19								
NUMBER_OF_PROCESSORS	4								
OS	Windows_NT								
Path	C:\Program Files\Common Files\Oracle\Java\javapath;C:\WIN								

Local Disk (C:)
 WinREAgent

> dell

cygwin64

MSOCache
 PerfLogs
 Program Files
 7-Zip
 Adobe
 Crashpad

CrowdStrike

Internet Explorer

> JAGS

> 📕 AMD

>

7. Give "JAVA\_HOME" as the variable name and the path to the JAVA program file as the Variable value. My default installation location is

"C:\Program Files\JAVA\jdk-21"

New System Variable			×				
Variable name: Variable value:	JAVA_HOME C:\Program Files\Java\jdk-21						
Browse Directory	. Browse File	ОК	Cancel				

Click 'OK', then click 'OK' on Environment Variables page and finally click 'OK' on System Properties page.

### Download and Install Anaconda 3

 Point your browser to <u>https://www.anaconda.com/download</u> and select "Download" button.



2. When the file

Anaconda3-2023.09-0-Windowsx86\_64.exe 1.0 GB • Done

#### is downloaded click to open and install

### Download and Install Anaconda 3

3. Find the Anaconda 3 folder in your list of apps and open the 'Anaconda' Navigator' (This may take a few minutes to launch)

Home	All applications 🗸 ON	base (root)
Environments	¢ DS	$\bigcirc$
	DataSpell	Anaconda Notebooks
Learning	DataSpell is an IDE for exploratory data analysis and prototyping machine learning models. It combines the interactivity of Jupyter notebooks with the intelligent	Cloud-hosted notebook service from Anaconda. Launch a preconfigured environment with hundreds of packag and store project files with persisten
Community	Python and R coding assistance of PyCharm in one user-friendly environment.	cloud storage.
	\$	2-6
	IP[y]:	100 A
	Qt Console	Spyder
	5.4.2 PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.	5.4.3 Scientific PYthon Development EnviRonment. Powerful Python IDE w advanced editing, interactive testing debugging and introspection feature
	(Lunat)	

4. Open the Spyder IDE by clicking "Launch" on that tile in the navigator (this may take a few minutes)

### Very Brief Orientation to Spyder IDE

😵 Spyder (Python 3.11)	- • ×
File Edit Search Source Run Debug Consoles Projects Tools View Help	
🗈 🖻 🖻 🕨 🗔 🗔 I, 🕨 🖓 🐺 🔹 🗉 🖋 🏓	C:\Users\mander10 - 🗁 🛧
C:\Users\mander10\untitled1.py	± 🗈 🕏 🔋 ९ ८ 🚍
temp.py X untitled0.py* X untitled1.py X	Name Type Size Value
1 # -*- coding: utf-8 -*-	df DataFrame (2, 8) Column names: Client(Deidentified), ClientAge atDOS(currentDOB),
3 Created on Thu Dec 14 13:43:53 2023	dfs list 1 [Dataframe]
5 @author: MANDER10	
7	
8	
	Created objects
Python Code File	
	Help Variable Explorer Plots Files
	Console 1/A ×
	In [9]:
	Console window
	IPython Console History

## In the console type 'pip install tabula-py'

### 1. Type 'pip install tabula-py' in the console window in the lower right

🕅 Spyde	er (Pyth	on 3.11)								 						٥	×
File Edit			Run Debi	ig Cons		jects Tools	s View I	Help									
	Ē	•			,,	₽ I⊳						¥	ę		C:\Users\mander10	-	t
C:\Users\	C:\Users\mander10\untitled0.py														± = = = = = = = = = = = = = = = = = = =		=
	emp.py 🕽	× untitle	d0.py 🗙											≡	Name Type Size Value		
1 2 3 4 5 6 7 <b>8</b>	# _* Crea @aut """	- codin ted on hor: MA	g: utf-8 Thu Dec NDER10	-*- 14 13:	:24:24	2023											
															Help Variable Explorer Plots Files	•	•
															Python 3.11.4   packaged by Anaconda, Inc.   (main, Jul 5 2023, 13:38:37) [MSC v.1916 64 bi Type "copyright", "credits" or "license" for more information. TP; mon 8.12.0 An enhanced Interactive Python. In [1]: pip install tabula-py	t (AME	064)]

## Run python commands in Spyder

### 1. Type the following commands into the **python code file**

import tabula import pandas as pd

dfs = tabula.read\_pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf",pages='all')
df = pd.concat(dfs, ignore\_index=True)

print(df)

Change to the location and name of your pdf

df.to\_csv("C:\\Users\\mander10\\Documents\\Medicaid variables2.csv", index=False)

Change to the location and name of desired csv



## Check the exported .csv

### • Here is what my exported csv looks like

6	5		▼ Medicaid variables2.csv - Excel											囨	-						
Fi	le	Home	Insert Page Layout Formulas Dat				a Review View JMP Acrobat ${igsilon}$ Tell me what you want to do											An	Anderson, Michael P.		
	<b>.</b> .	Calibr	i ·	11 × Â Â		≫	📴 Wrap	o Text	(	General	Ŧ				<b>+</b> =		Σ	Ž	2	E.	
Past	e 💉	B <i>I</i>	<u>U</u> -	- <u>8</u> - <u>A</u>	===	€ →	🖶 Mero	ge & Cent	er *	\$-%,	€.0 .00 .00 <b>→</b> .0	Conditional Formatting *	Format as Table *	Cell Styles *	Insert *	Delete Fo	rmat	Sort 8 Filter	& Find & ▼ Select ▼	Create PDF and Share link S	C Sh
Clipt	board	Gi I	Fo	nt i	ā l	Aligr	nment		Es.	Number	E.		Styles			Cells		Editi	ng	Adobe	A A
A1		<b>•</b>	×	√ f <sub>x</sub>	Client(Deide	ntified)															
		А			В			С			D		E		F		G		Н	I	
1	Client(	(Deidenti	fied)	ClientAge atDOS	(currentDOB	) Clie	ent Age(C	omplete	dMont	hs) Race(	Group	EthnicCode		S	exCode	e First	Dateof	Service	Provider	(Deidentified)	
2	C#		C	0-18			0-250			Race		Hispanicor Non-Hispanic		nic M	M or F xx/xx/xx		x/xxxx		P#		
3	B#		1	19-25		25:	251-470			Race	ace Hispanicor Non-Hispanic			nic M	M or F Yy/yy/yyyy			D#			
4																					

- Looks good for the most part but looking closely we see some of the spaces are missing in both the variable names and in that data.
- PDF often leave a residual '\r' that causes this. Let's remove it.

## Rerun python commands in Spyder

1. Type the following commands into the **python code file** 

import tabula import pandas as pd dfs = tabula.read\_pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf",pages='all') df = pd.concat(dfs, ignore\_index=True) df.columns = df.columns.str.replace('[\r]', ' ') df.replace('\r', ' ', regex=True, inplace=True)

Replace \r with a space in the data

print(df)

df.to\_csv("C:\\Users\\mander10\\Documents\\Medicaid variables2.csv", index=False)

Change to the location and name of desired csv

### Rerun python commands in Spyder

• Code with formatting to remove '\r' is below

```
# -*- coding: utf-8 -*-
.....
Created on Thu Dec 14 13:24:24 2023
@author: MANDER10
.....
import tabula
import pandas as pd
dfs = tabula.read pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf",pages='all')
df = pd.concat(dfs, ignore index=True)
print(df)
df.columns = df.columns.str.replace(' [ \ r ]', ' ')
df.replace('\r', '', regex=True, inplace=True)
df.to csv("C:\\Users\\mander10\\Documents\\Medicaid variables2.csv", index=False)
```

## Check the exported .csv

#### • Here is what my exported csv looks like now

6	ר <i>י</i> יר ∎	-				Medicaid v		Ē		$\times$			
F	File Home	Insert Page Layout	Formulas D	ata Review	View JMP	Acrobat 🛛 🖓		Anderson, Michael P. (HSC) $$ $$ Share					
Pas	Calibri ste	• 11 • A /		≫ -	Text	ieneral	Conditional Format as	Cell Inse	ert Delete Format	$\Sigma \stackrel{\bullet}{} A_{\overline{Z}} \qquad $	Create PDF	Create PDF and	
Clip	board 😨	Font	G	Alignment	Fa	Number	Formatting Table Sites	tyles •	Cells	Filter - Select - Editing	and Share link S Adobe	hare via Outlook Acrobat	~
E1	.0 -	$\times \checkmark f_x$											^
	A	E	3		С	D	E	F	G	Н	I J	К	
1	Client (Deidentif	ied) Client Age at DO	S (current DOB)	Client Age (Con	npleted Month	s) Race Group	Ethnic Code	Sex Code	First Date of Serv	vice Provider (De	identified)		
2	C#	0-18		0-250		Race	Hispanic or Non-Hispanic	M or F	xx/xx/xxxx	P#			
3	B#	19-25		251-470		Race	Race Hispanic or Non- Hispanic		Υγ/γγ/γγγγ	D#			
4													

• Spaces are now corrected for both the column headings and the data