

Extracting Data from PDF to CSV format using Python

Michael Anderson, PhD

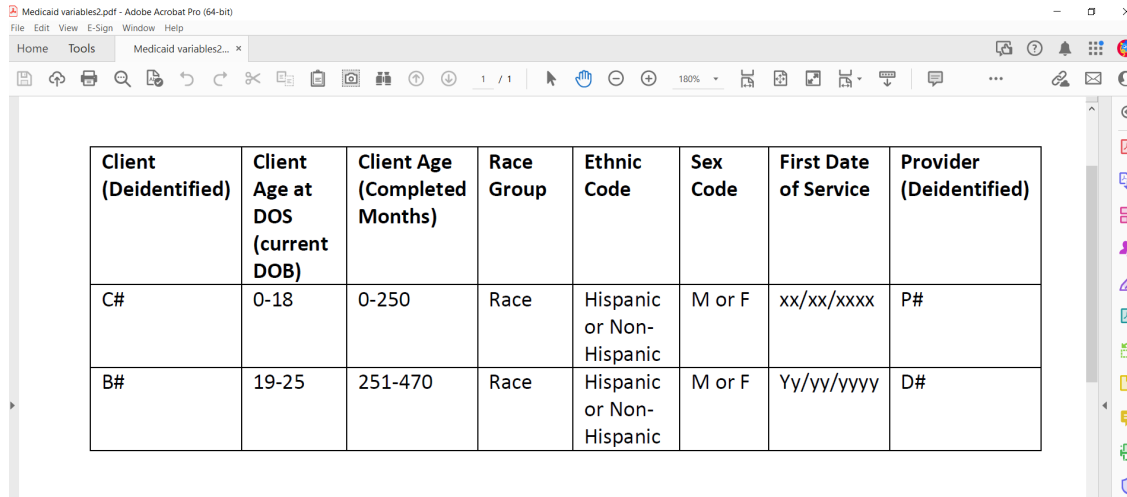
Director, Design and Computational Resources Unit

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.



The screenshot shows a PDF document titled "Medicaid variables2.pdf" in Adobe Acrobat Pro. The document contains a table with the following structure:

Client (Deidentified)	Client Age at DOS (current DOB)	Client Age (Completed Months)	Race Group	Ethnic Code	Sex Code	First Date of Service	Provider (Deidentified)
C#	0-18	0-250	Race	Hispanic or Non-Hispanic	M or F	xx/xx/xxxx	P#
B#	19-25	251-470	Race	Hispanic or Non-Hispanic	M or F	Yy/yy/yyyy	D#

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

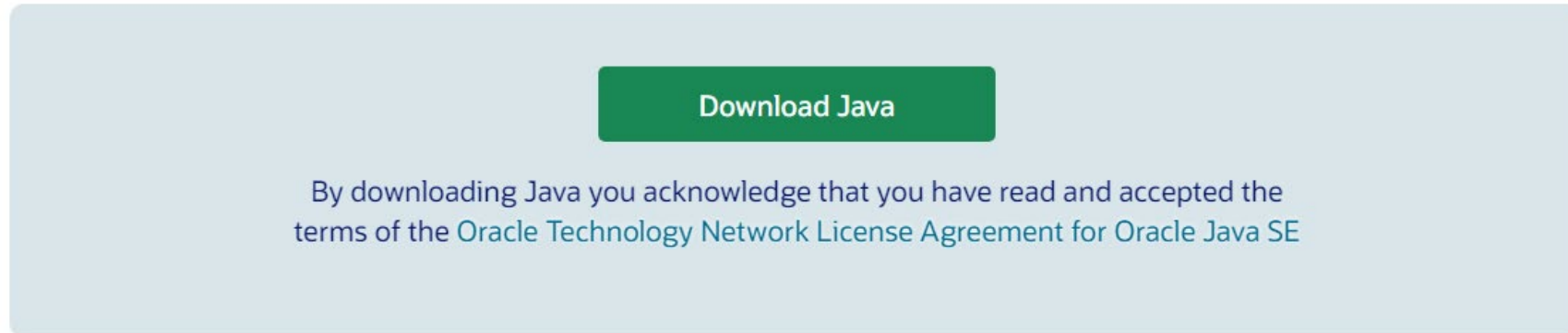
Subsequent Use

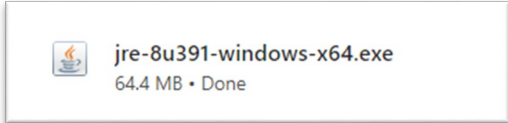
1. Run python commands in Spyder
2. Check the exported .csv and adjust Python commands as needed.

Let's look at each of these steps in detail

Download and Install Java (admin approval needed)

1. Navigate to <https://www.java.com/en/download/> and select “Download Java”

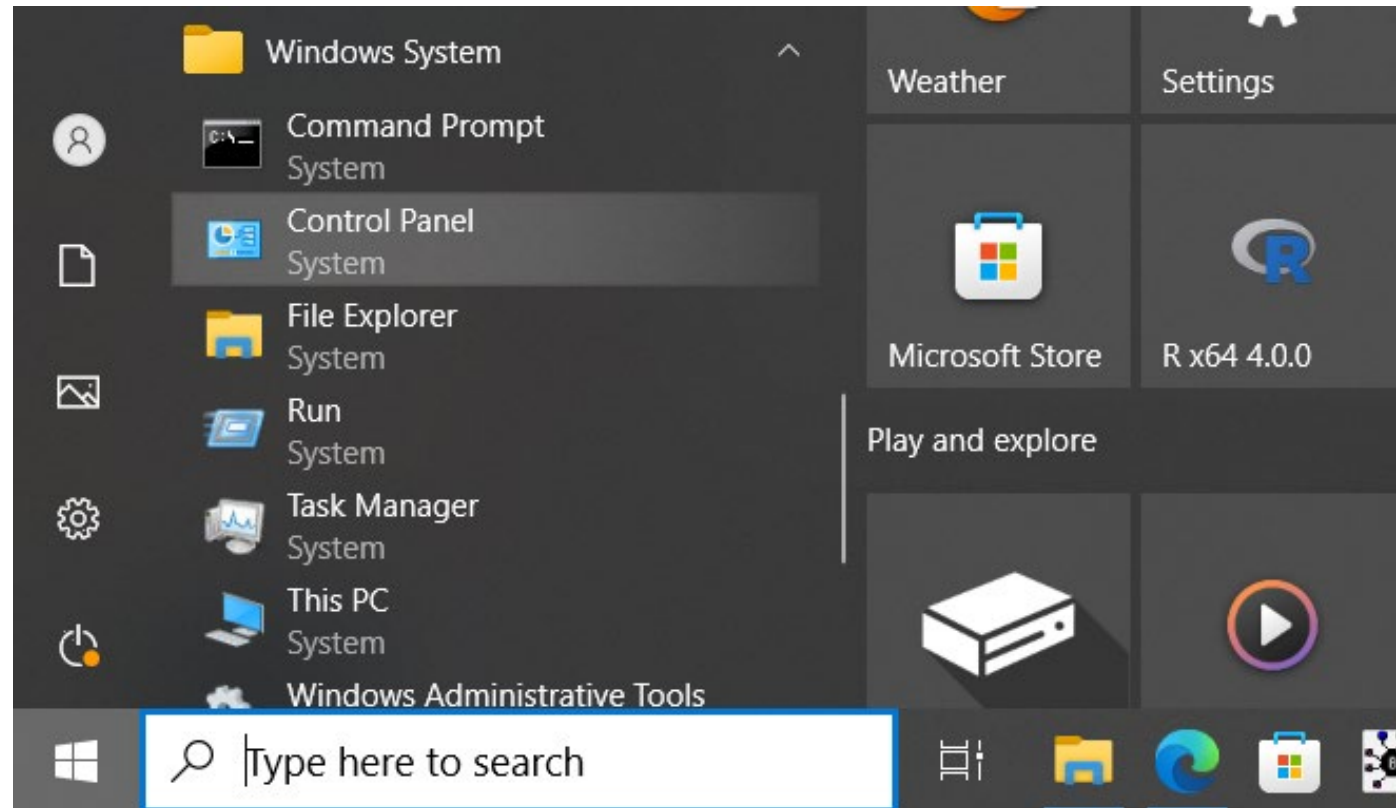


2. When the file  is downloaded, click to open and install

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.

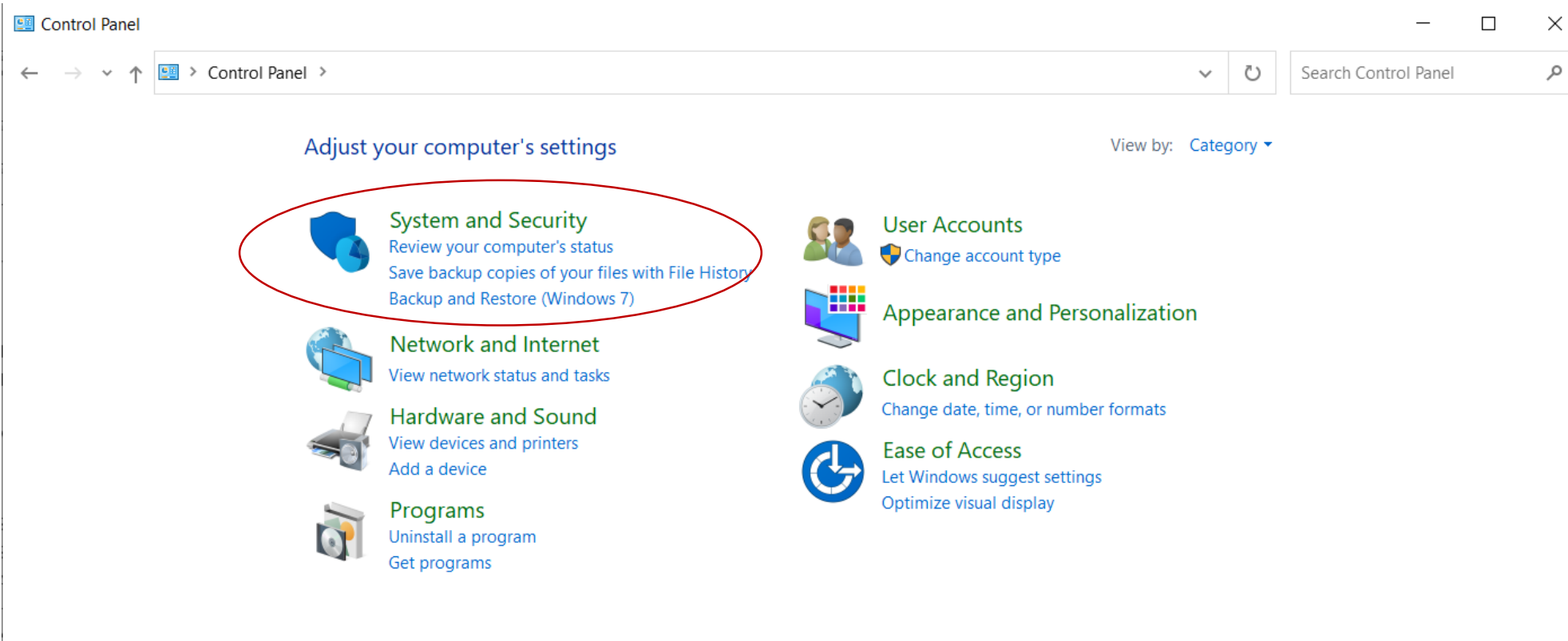
Create new JAVA_HOME environment variable (admin approval needed)

1. Launch “Control Panel” from Windows System folder



Create new JAVA_HOME environment variable (admin approval needed)


2. Select System and Security



Create new JAVA_HOME environment variable (admin approval needed)



3. Select System


System and Security



← → ▾ ↑  > Control Panel > System and Security > ▾ ↻


Control Panel Home

- **System and Security**
- Network and Internet
- Hardware and Sound
- Programs
- User Accounts
- Appearance and Personalization
- Clock and Region
- Ease of Access

 **Security and Maintenance**
Review your computer's status and resolve issues |  Change User Account Control settings |
Troubleshoot common computer problems

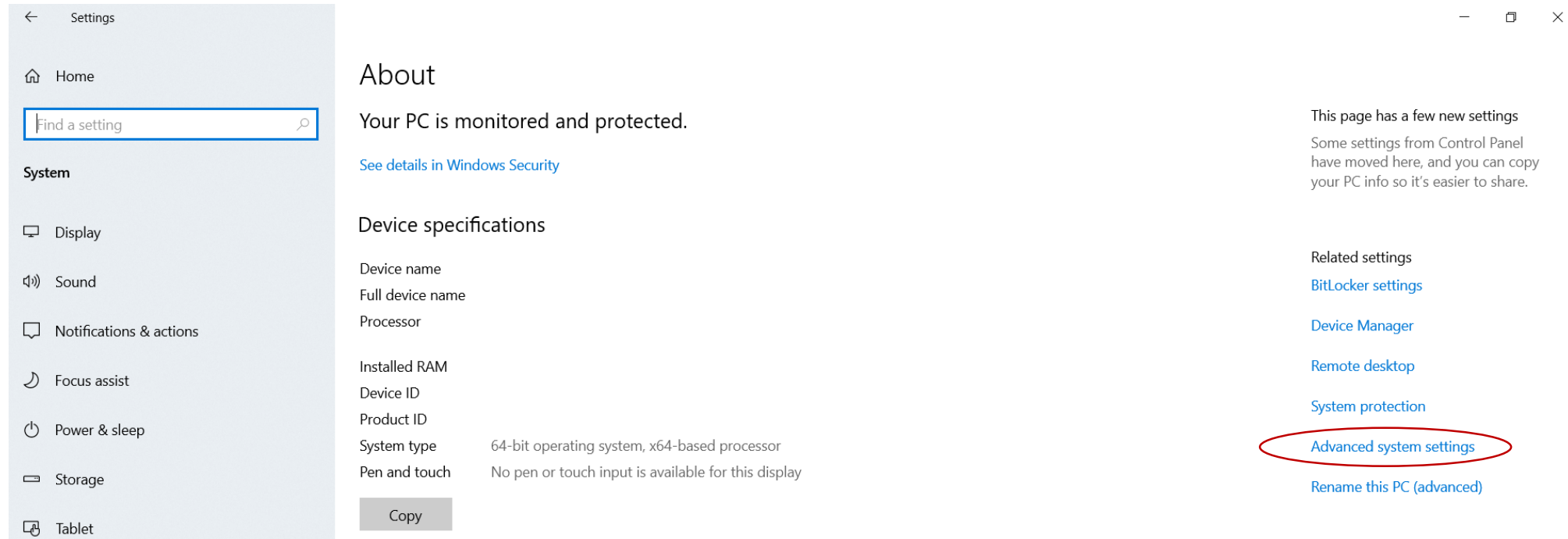
 **Windows Defender Firewall**
[Check firewall status](#) | [Allow an app through Windows Firewall](#)

 **System**
[View amount of RAM and processor speed](#) |  Allow remote access | [Launch remote assistance](#) |
[See the name of this computer](#)

 **Power Options**
[Change what the power buttons do](#) | [Change when the computer sleeps](#)

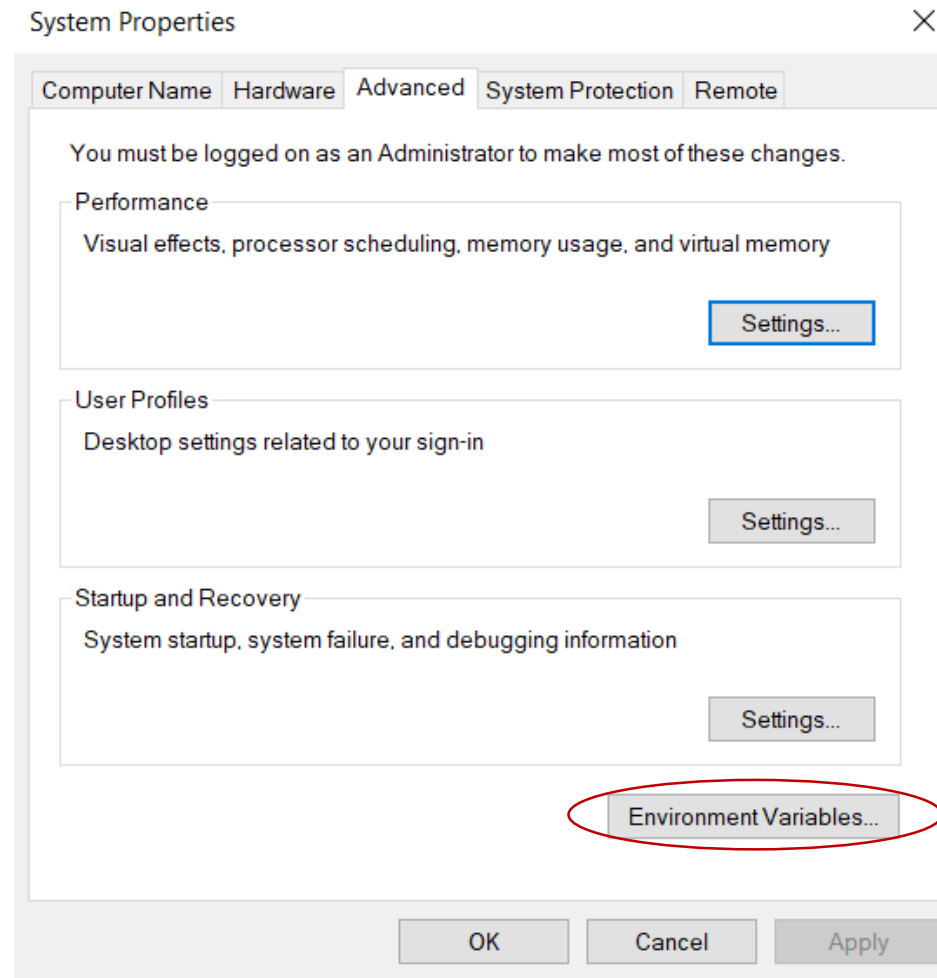
Create new JAVA_HOME environment variable (admin approval needed)

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



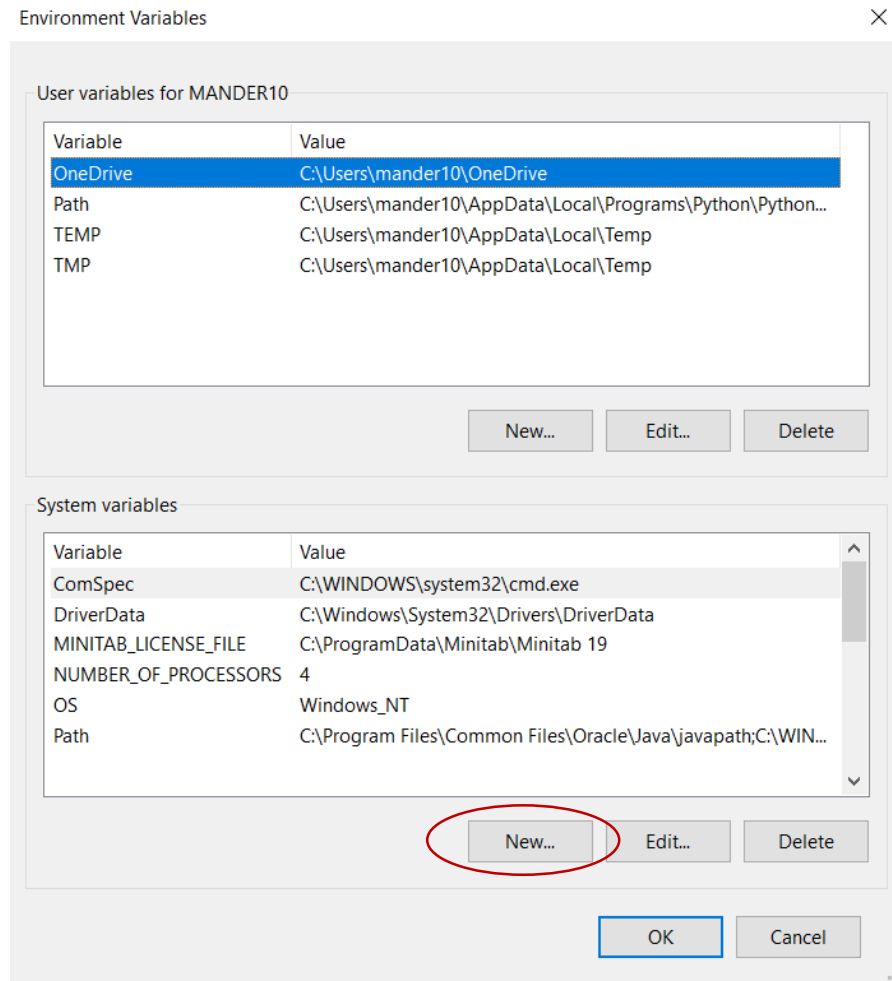
Create new JAVA_HOME environment variable (admin approval needed)

5. Select “Environment Variables” button



Create new JAVA_HOME environment variable (admin approval needed)

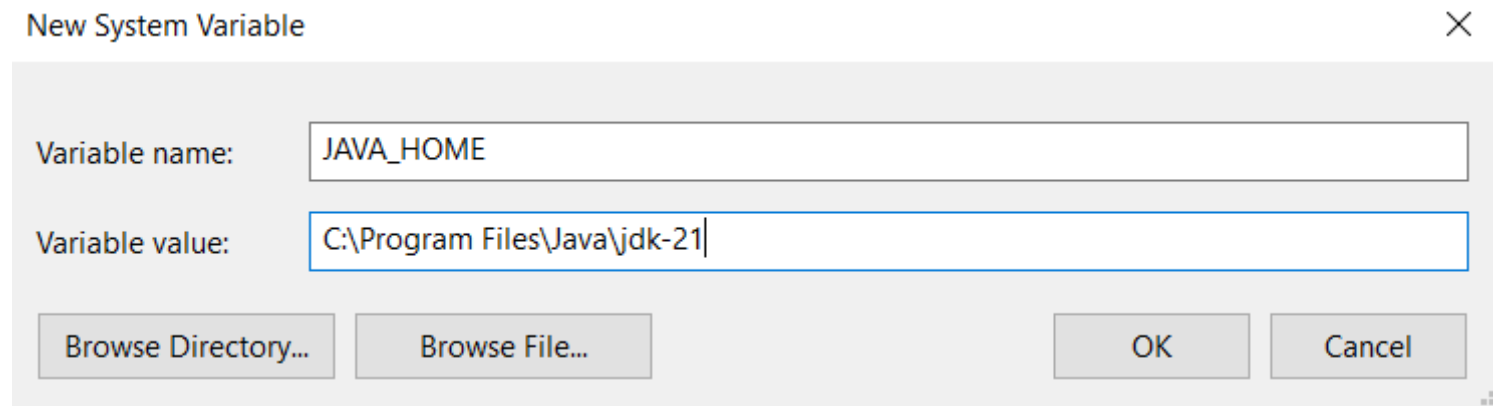
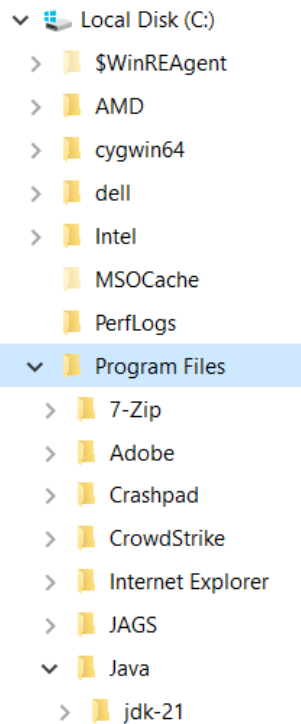
6. Select “New” from the System variables box



Create new JAVA_HOME environment variable (admin approval needed)

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

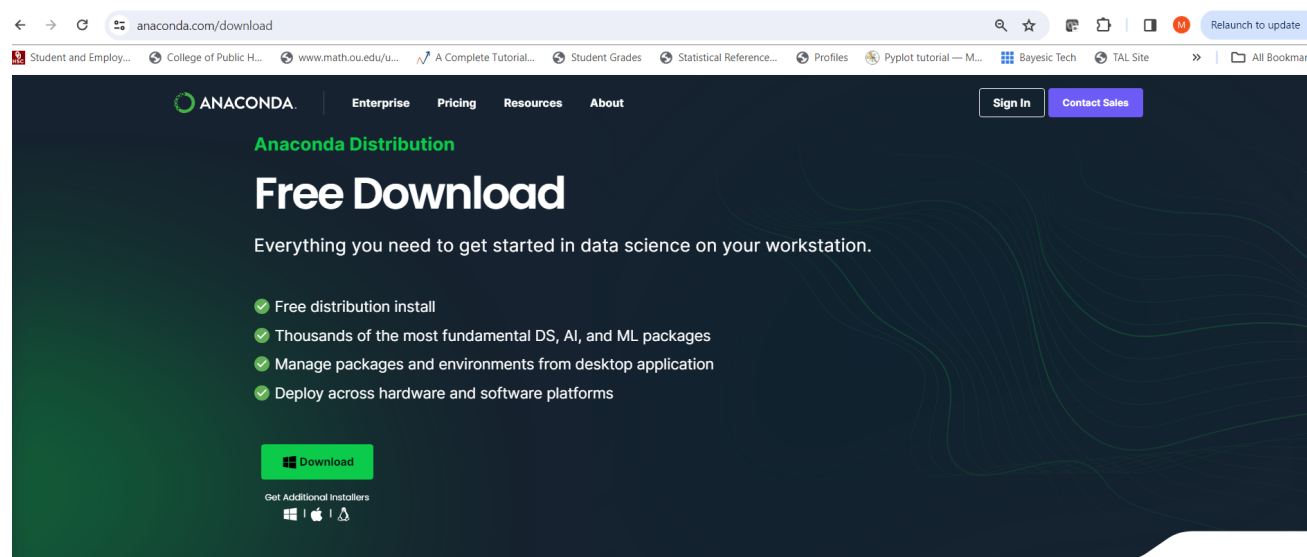
“C:\Program Files\JAVA\jdk-21”



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

Download and Install Anaconda 3

1. Point your browser to <https://www.anaconda.com/download> and select “Download” button.



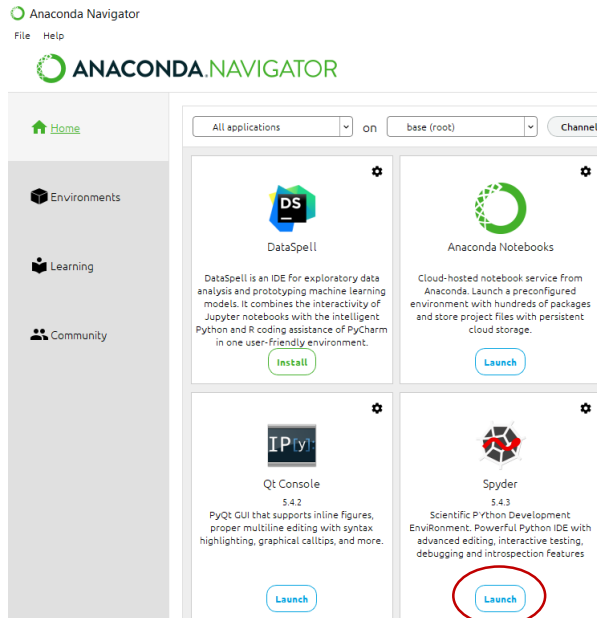
2. When the file

Anaconda3-2023.09-0-Windows-x86_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)



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

The image shows the Spyder IDE interface with three key components highlighted by white ovals:

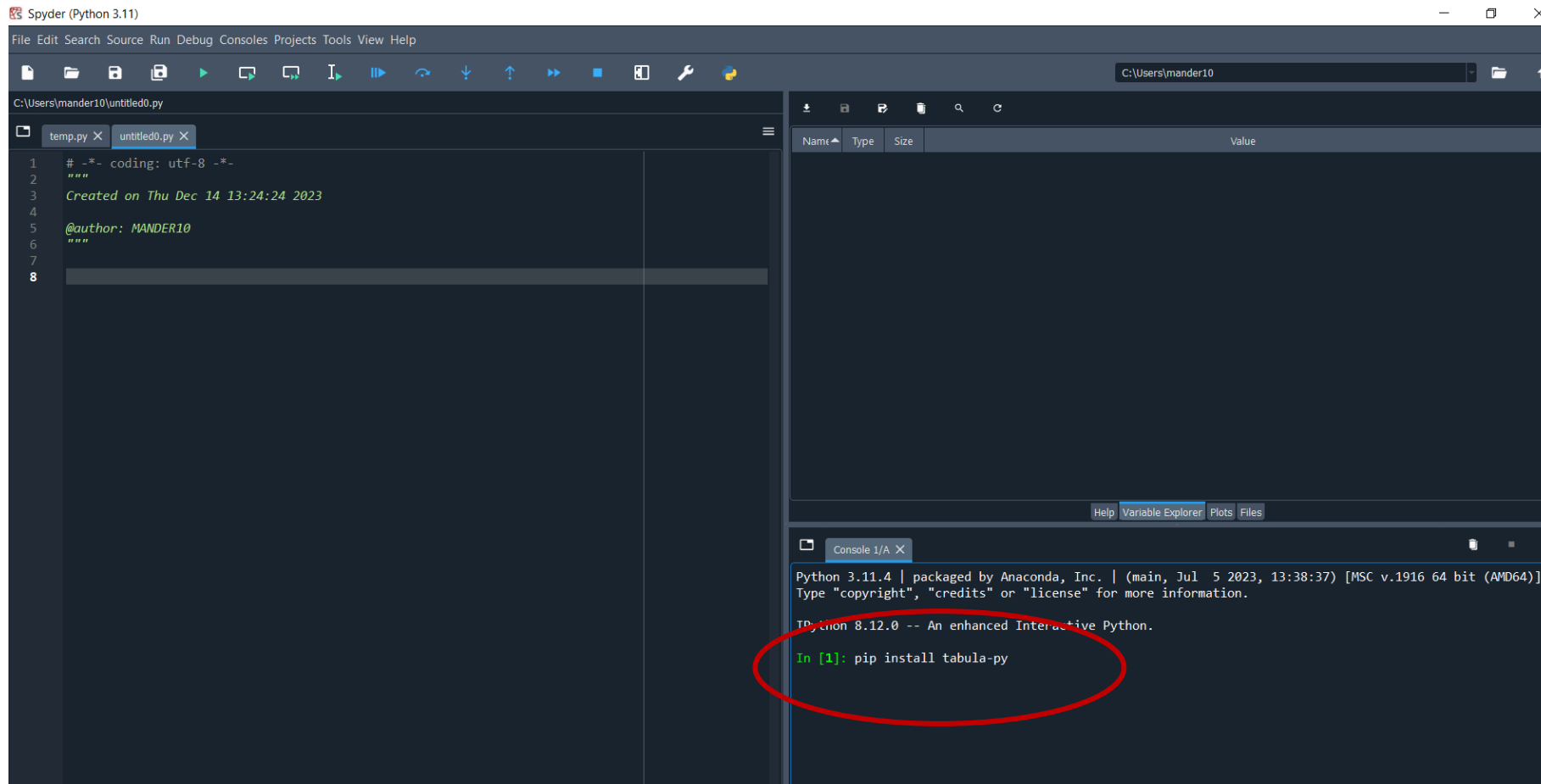
- Python Code File:** The left pane shows a Python script with a docstring and a timestamp:

```
1 # -*- coding: utf-8 -*-  
2 """  
3 Created on Thu Dec 14 13:43:53 2023  
4  
5 @author: MANDER10  
6 """  
7  
8
```
- Created objects:** The top-right pane displays a table of objects in memory:

Name	Type	Size	Value
df	DataFrame	(2, 8)	Column names: Client(Deidentified), ClientAge atDOS(currentDOB), ...
dfs	list	1	[Dataframe]
- Console Window:** The bottom-right pane shows the IPython console with the prompt `In [9]:`.

In the console type 'pip install tabula-py'

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



```
Spyder (Python 3.11)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\mander10\untitled0.py
temp.py x untitled0.py x
1 #-*- coding: utf-8 -*-
2 """
3 Created on Thu Dec 14 13:24:24 2023
4
5 @author: MANDER10
6 """
7
8
C:\Users\mander10
Name Type Size Value
Help Variable Explorer Plots Files
Console 1/A x
Python 3.11.4 | packaged by Anaconda, Inc. | (main, Jul 5 2023, 13:38:37) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.
Python 8.12.0 -- An enhanced Interactive Python.
In [1]: pip install tabula-py
```


Run python commands in Spyder

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

```
import tabula
```

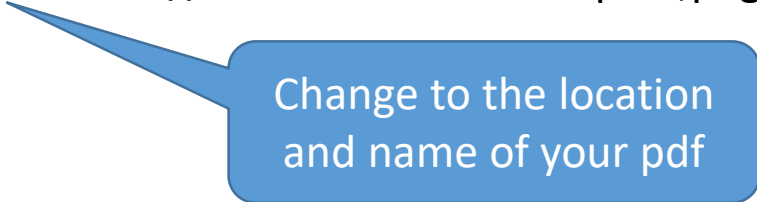
```
import pandas as pd
```

```
dfs = tabula.read_pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf",pages='all')
```

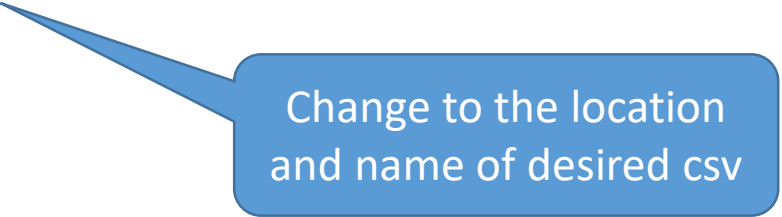
```
df = pd.concat(dfs, ignore_index=True)
```

```
print(df)
```

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

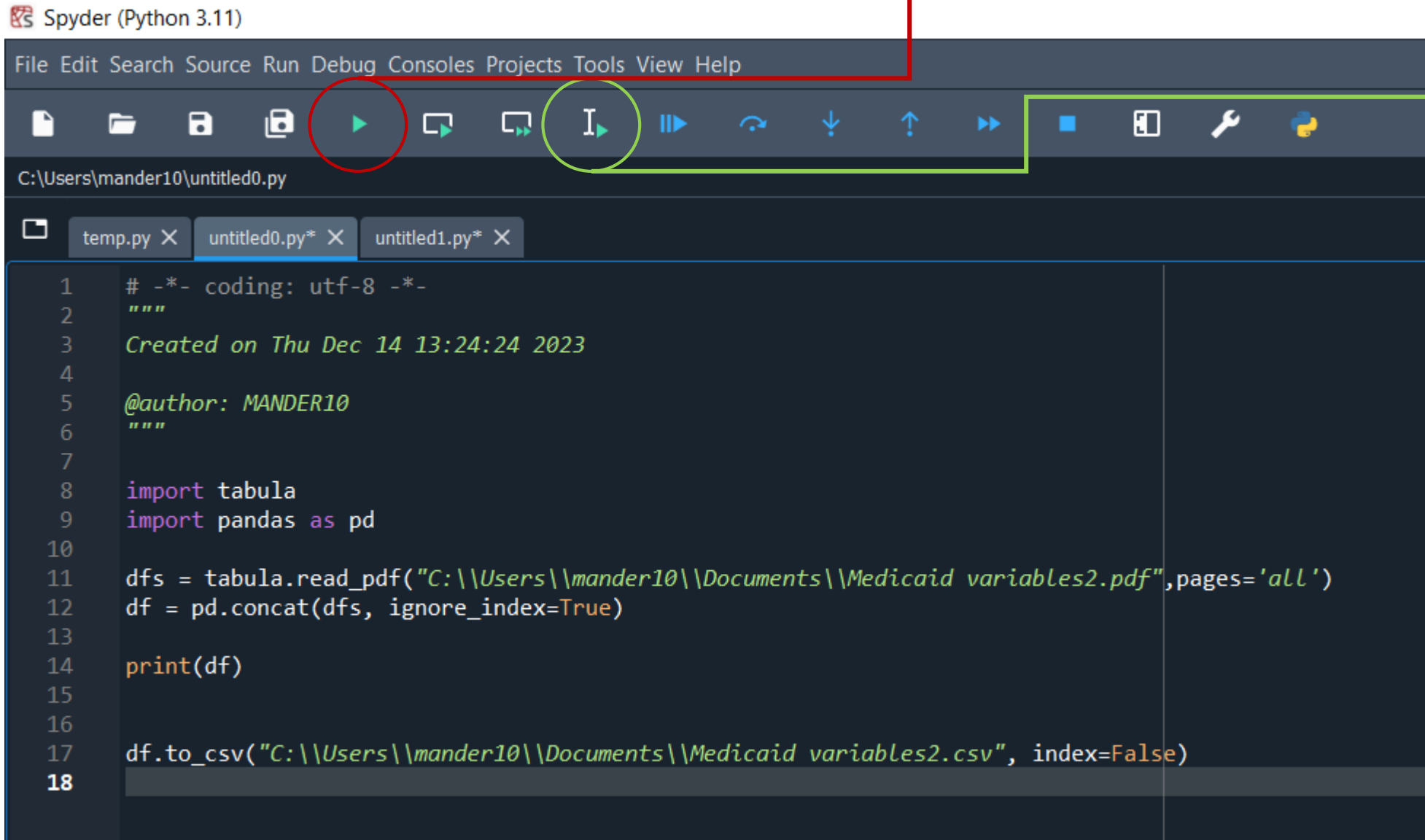


Change to the location
and name of your pdf



Change to the location
and name of desired csv

Run python commands in Spyder



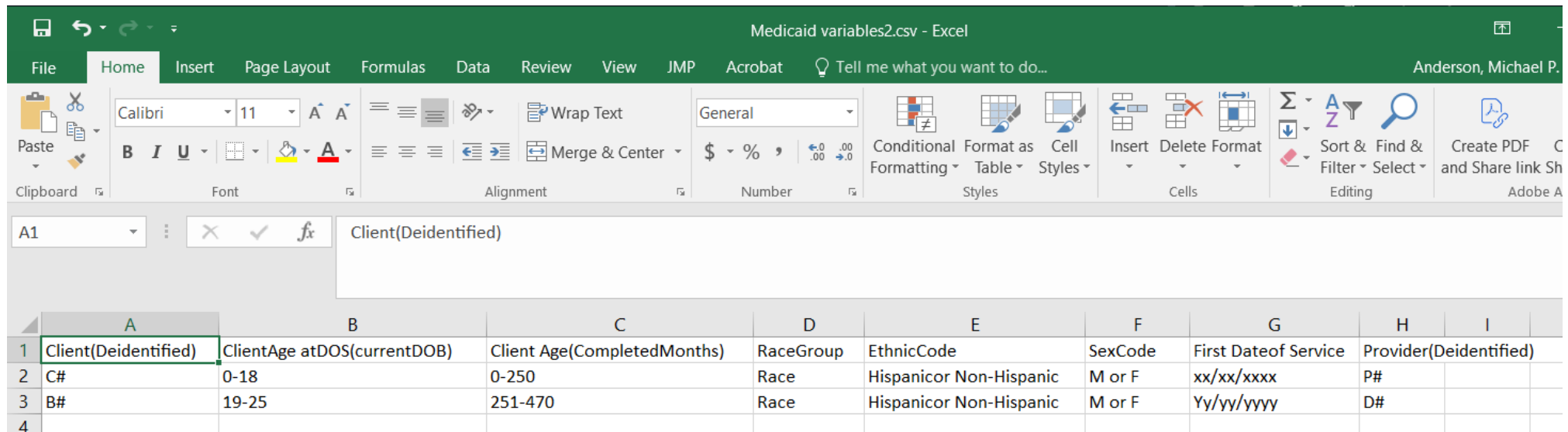
Run all code

Run current line

Run all the code.
This should create
the csv.

Check the exported .csv

- Here is what my exported csv looks like



	A	B	C	D	E	F	G	H	I
1	Client(Deidentified)	ClientAge atDOS(currentDOB)	Client Age(CompletedMonths)	RaceGroup	EthnicCode	SexCode	First Dateof Service	Provider(Deidentified)	
2	C#	0-18	0-250	Race	Hispanicor Non-Hispanic	M or F	xx/xx/xxxx	P#	
3	B#	19-25	251-470	Race	Hispanicor Non-Hispanic	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
```

Change to the location
and name of your pdf

```
dfs = tabula.read_pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf",pages='all')  
df = pd.concat(dfs, ignore_index=True)
```

Replace \r with a space in the column headers

```
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

```
1  # -*- coding: utf-8 -*-
2  """
3  Created on Thu Dec 14 13:24:24 2023
4
5  @author: MANDER10
6  """
7
8  import tabula
9  import pandas as pd
10
11  dfs = tabula.read_pdf("C:\\Users\\mander10\\Documents\\Medicaid variables2.pdf", pages='all')
12  df = pd.concat(dfs, ignore_index=True)
13
14  print(df)
15
16  df.columns = df.columns.str.replace('[\r]', ' ')
17  df.replace('\r', ' ', regex=True, inplace=True)
18
19  df.to_csv("C:\\Users\\mander10\\Documents\\Medicaid variables2.csv", index=False)
20
```

Check the exported .csv

- Here is what my exported csv looks like now

The screenshot shows the Microsoft Excel interface with the following data:

	A	B	C	D	E	F	G	H	I	J	K
1	Client (Deidentified)	Client Age at DOS (current DOB)	Client Age (Completed Months)	Race Group	Ethnic Code	Sex Code	First Date of Service	Provider (Deidentified)			
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	Hispanic or Non- Hispanic	M or F	Yy/yy/yyyy	D#			
4											

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