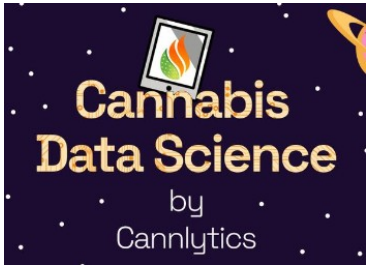
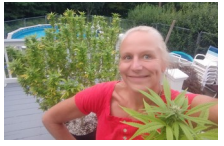


https://www.meetup.com/cannabis-data-science/

Cannlytics (Cannabis Analytics Free Open Source Software)
- web scraping with Selenium and Beautiful Soup
- parsing pdf files with PDFplumber

Candace O`Sullivan-Sutherland 8/27/2022

https://github.com/candy-o
[2021-present] Software Engineer/Data Scientist @ cannlytics.com
[2019-present] Software Engineer/Data Scientist @ TSSG
[1998-2018] Senior Software Performance Engineer @ Dell Technologies DellEMC (DG CLARiiON)
[1995-1998] Senior Software Engineer @ Digital Equipment Corporation, StorageWorks Division

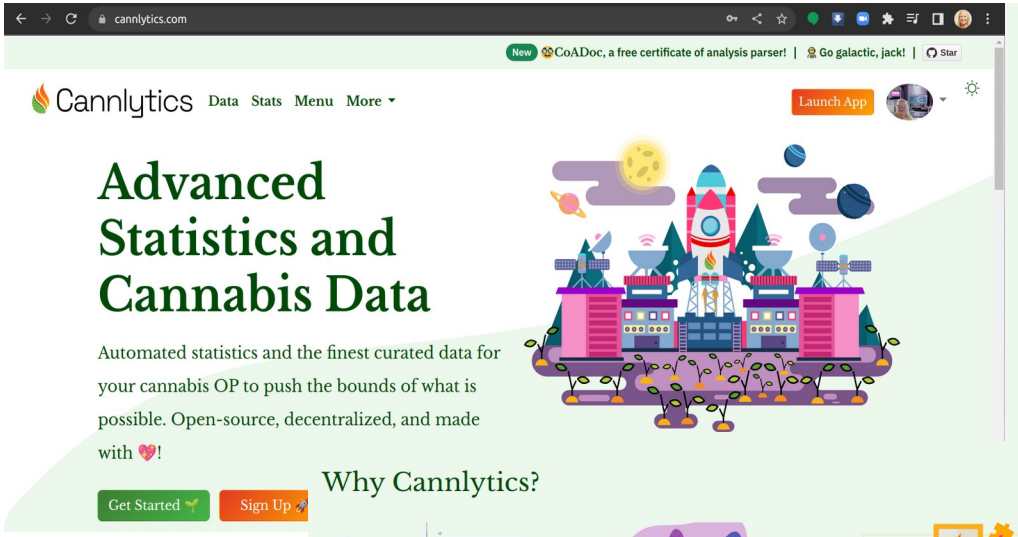


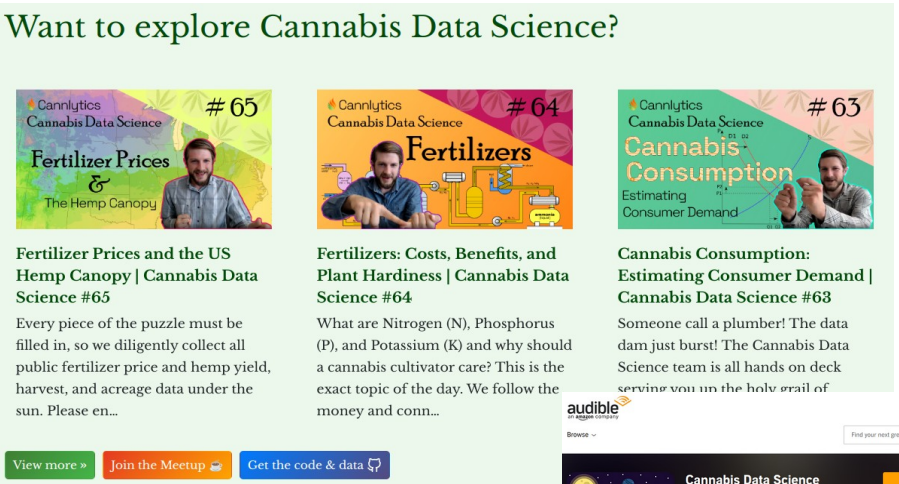
https://www.meetup.com/cannabis-data-science/
https://cannlytics.com
https://github.com/cannlytics

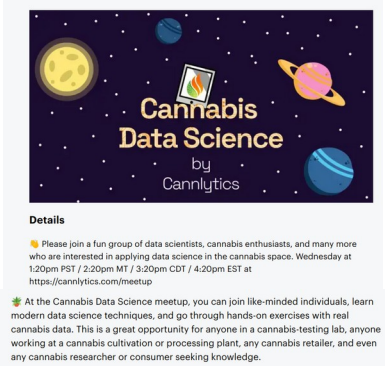
Created using LibreOffice Impress https://www.libreoffice.org/ (free software / open source software)

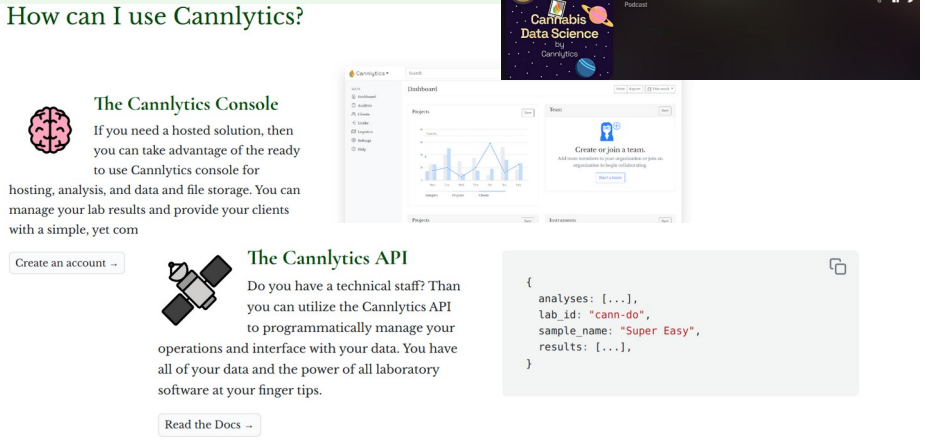
Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

“Cannlytics” is an Open-source, Decentralized Cannabis Data Science/Analytics organization offering a suite of free software for the Cannabis industry.









Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Cannlytics (creator “Keegan” and Cannlytics Team) built with ❤

cannlytics.com/team

Cannlytics is built with the love of...

These heros who create, make, and deliver your software solutions and support.

Keegan Skeate
KEO

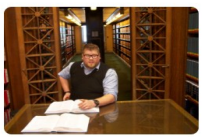
Keegan is an economist, software developer, and data scientist who cut his chops as a laboratory analyst. Keegan will work tirelessly to deliver you value for value.

Alice Allafort, Ph.D.
Data Scientist

Alice graduated with a Physics Ph.D. from Stanford University and is always happy to work on interesting data science projects.

Jeffrey Miller
Customer Outreach Developer

Jeff is a customer outreach specialist who revolves around creating value and lives and breathes customer satisfaction.

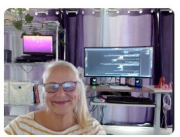


Raymond Bentson
International Communications Specialist / Linguist

Raymond is experienced in agriculture, international trade, and language services. Raymond is also the founder of Advanced Language Liaison.

Michael Pilosov, Ph.D.
Project Manager

Michael is an applied mathematician on a mission to democratize access to quantitative decision-making tools. Micahel is here to help you know your data inside and out and interface with your analytics to get the insights that you need.



Candace O'Sullivan-Sutherland
Data Scientist

Candace is a talented data and computer scientist with all the tools in her toolbelt to handle the biggest of big data. Candace brings an extensive suite of hardware capabilities to ensure that your data pipelines are pristine.

CannBot
Chief Scientific Robot

CannBot does the dirty work, working meticulously around the clock to automate the boring stuff and do the heavy lifting.

Cat the Caterpillar
Chief Organic Organism

Cat, the data-pipeline caterpillar, ensures that your data is clean, organized, and gets to where you need it.

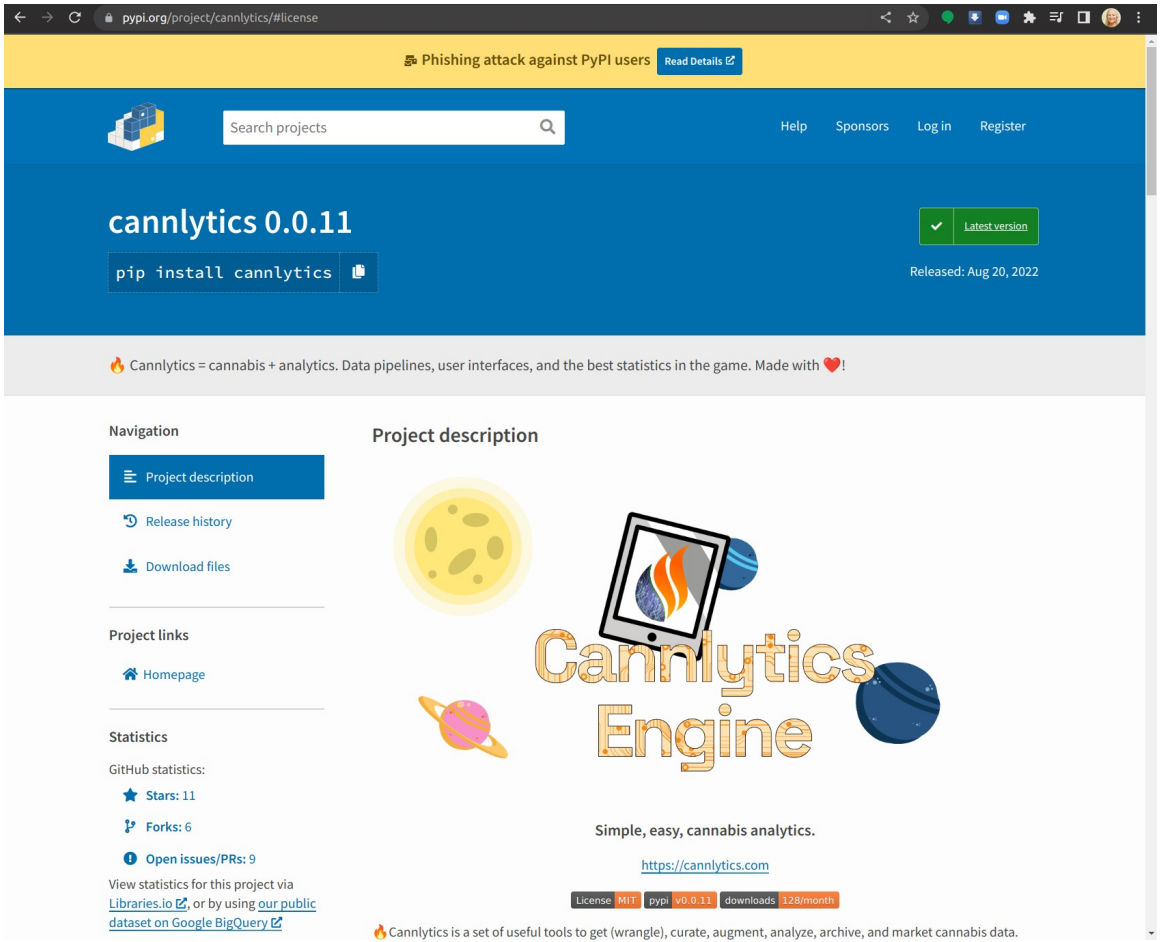
Interested in joining?

Do you have the chops and / or passion to help deliver industry-leading analytics, software, and support to everyone in the cannabis industry? Then please email the team at dev@cannlytics.com and explain who, what, where, why, when, and how you can contrbiute to Cannlytics.

Join today →

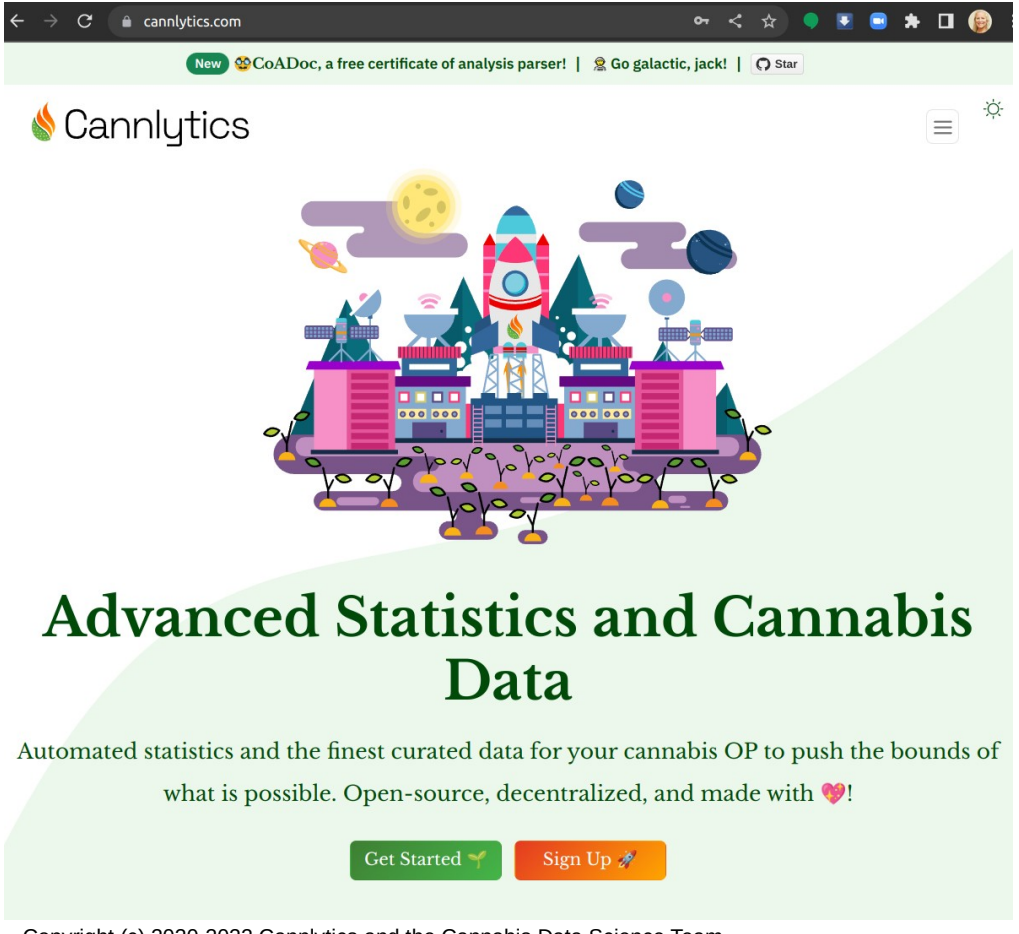
Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

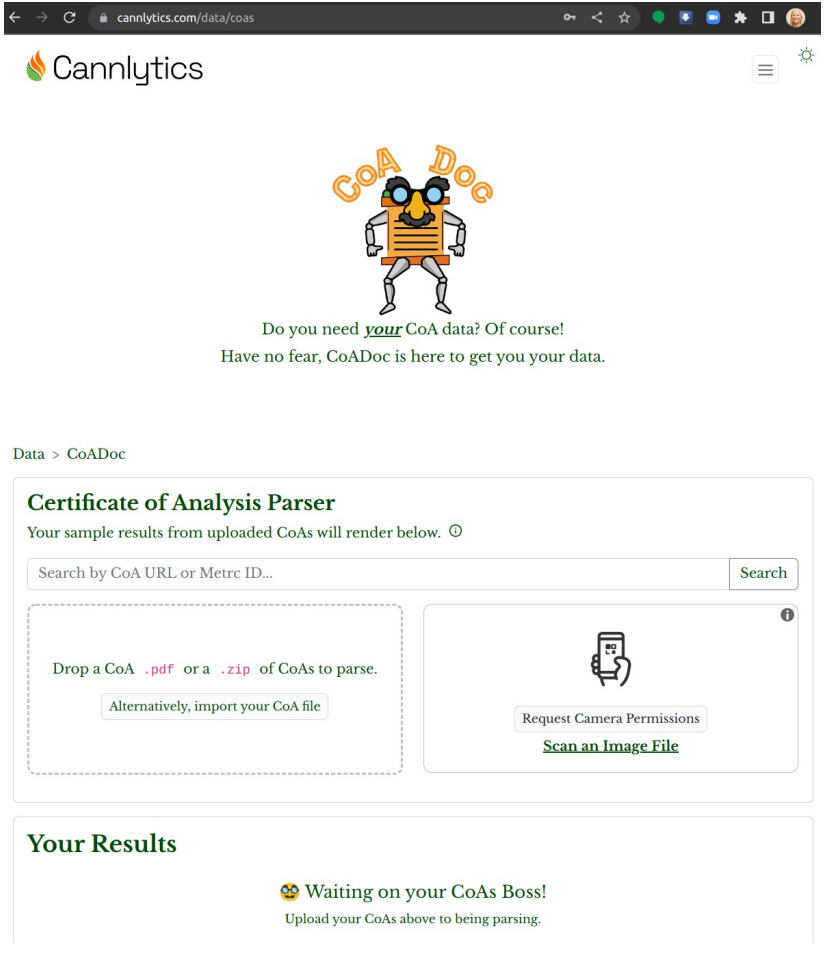
`cannlytics 0.0.11` pipy.org released Aug 20, 2022



Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

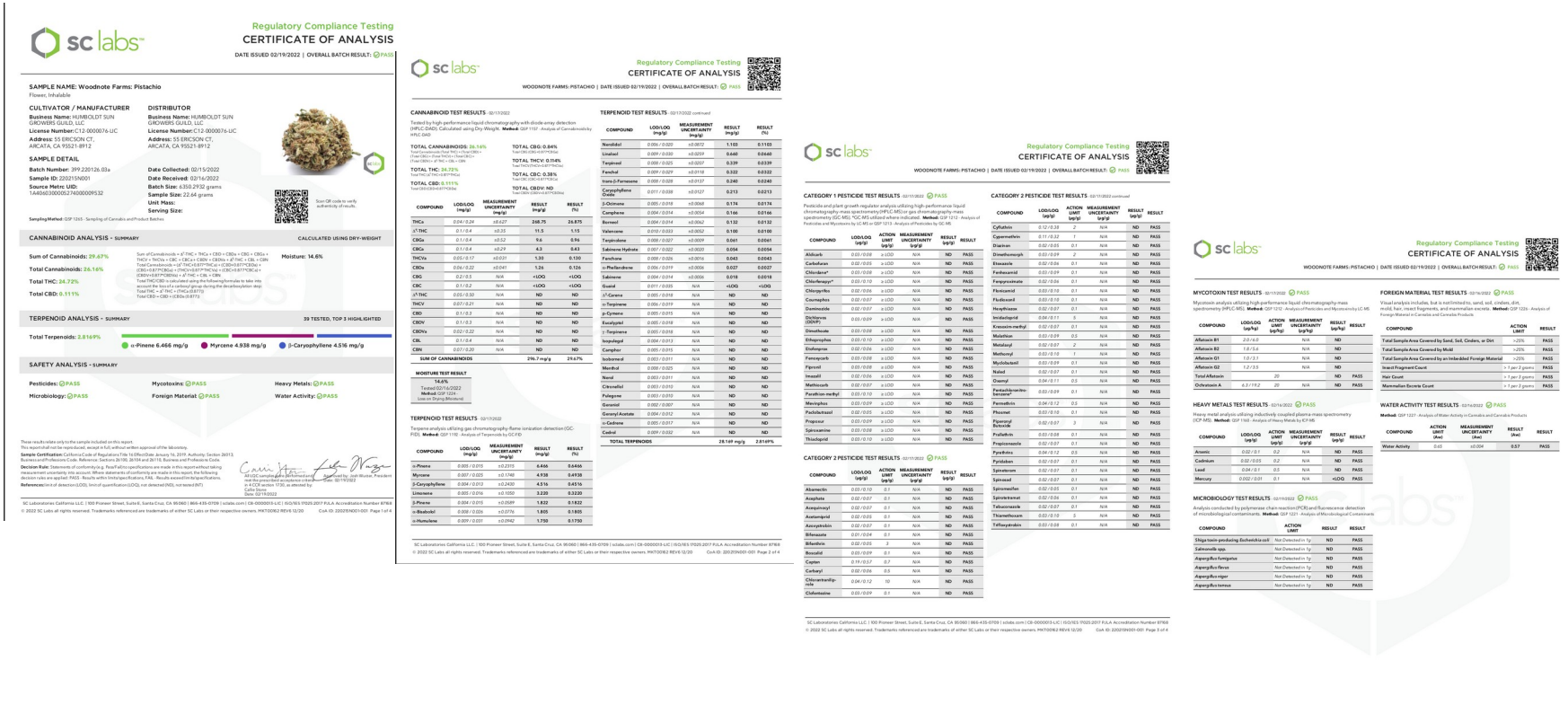
NEW Cannlytics engine CoADoc reads and parses COA Cannabis test lab “Certificates of Analysis” pdf file(s) and saves to .xlsx format.





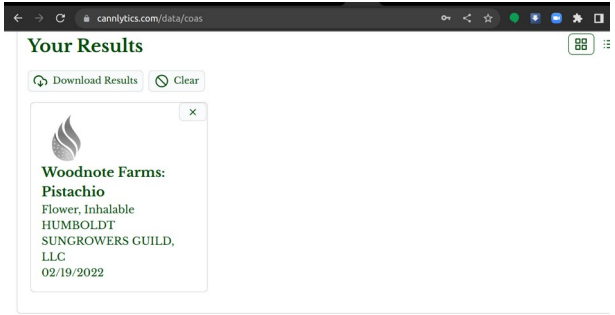
Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Lets drag this COA ‘pistachio.pdf’ (4pgs) onto CoADoc



Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

CoADoc read, parsed and placed our data into a table

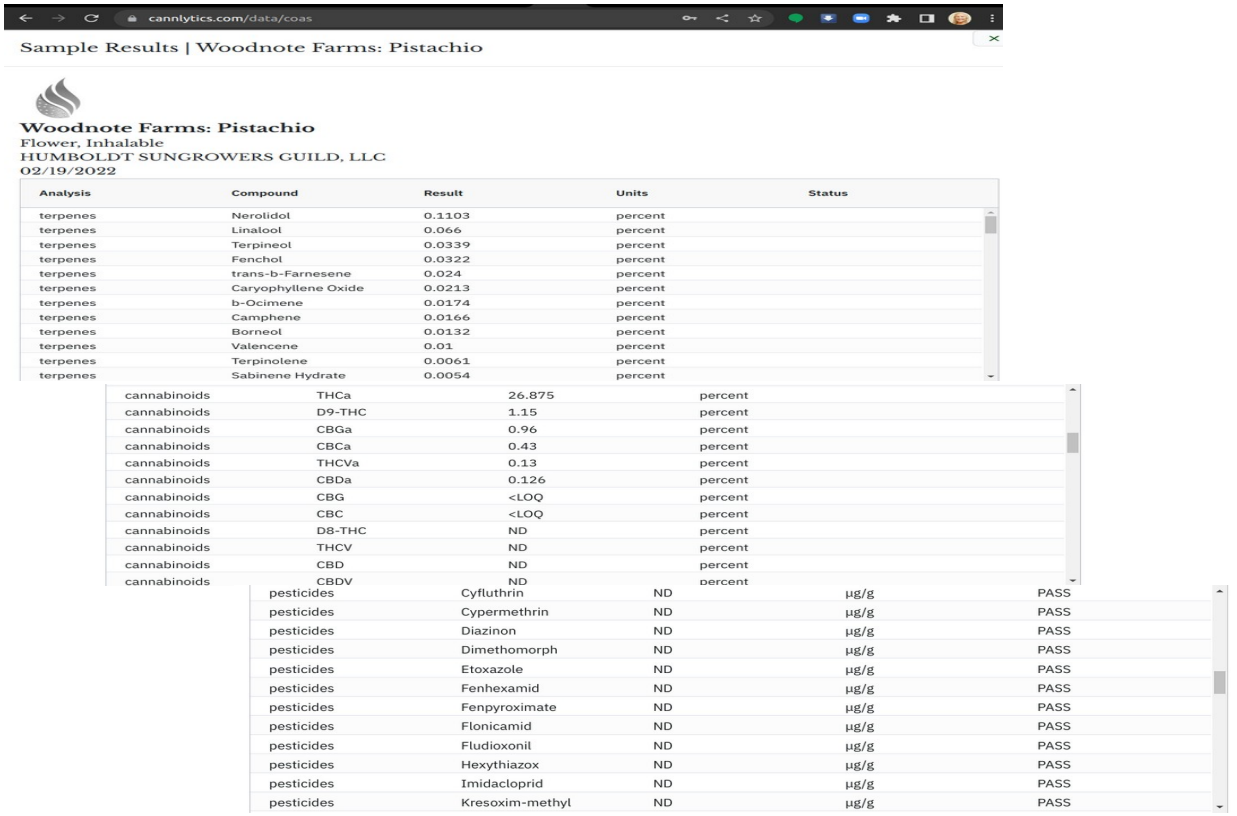


Custom CoA Parsing

At this time, CoADoc can only parse certificates of analysis (CoAs) from labs and LIMS with validated parsing algorithms. We've validated:

Labs	LIMS
Anresco Laboratories	Confident Cannabis
Cannalysis	TagLeaf LIMS
Green Leaf Lab	
MCR Labs	
SC Labs	
Sonoma Lab Works	
Veda Scientific	

If you want your favorite lab or LIMS added, then please email **dev@cannlytics.com** and chances are that they can be included. Alternatively, because

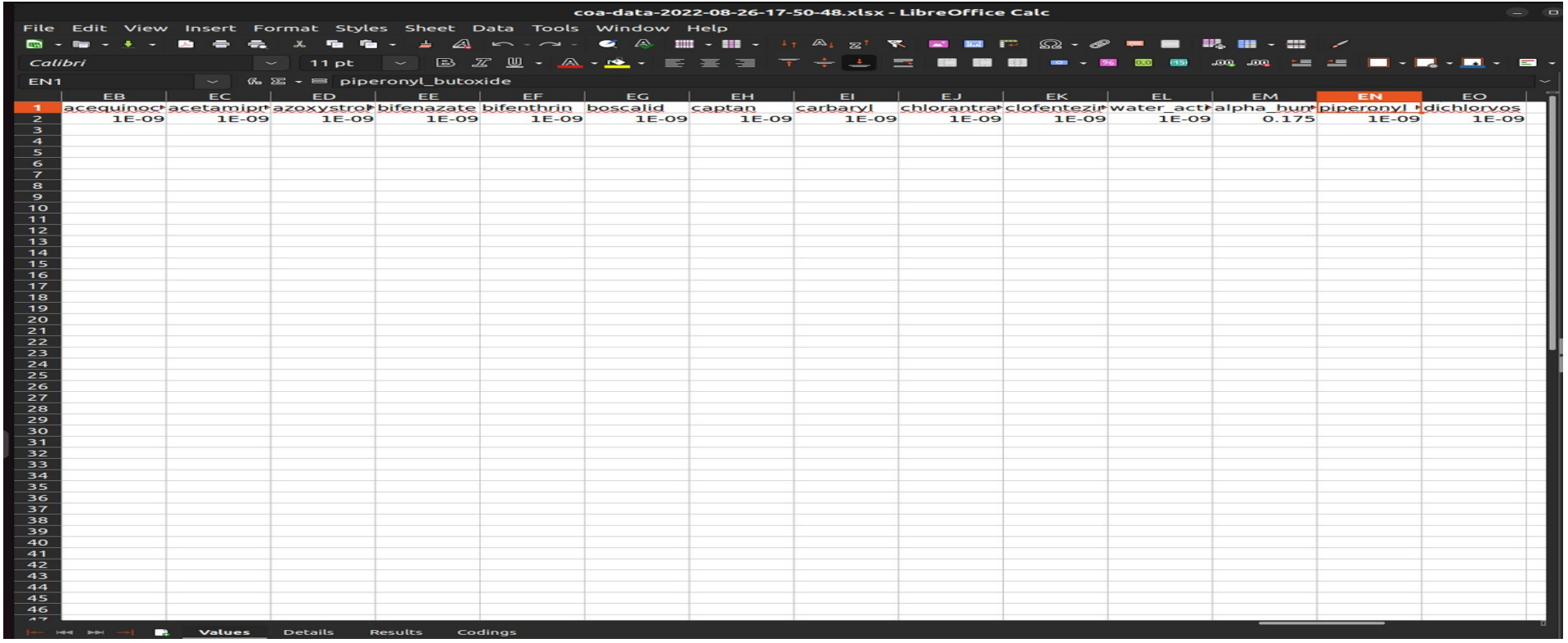


Analysis	Compound	Result	Units	Status
terpenes	Nerolidol	0.1103	percent	
terpenes	Linalool	0.066	percent	
terpenes	Terpineol	0.0339	percent	
terpenes	Fenchol	0.0322	percent	
terpenes	trans-b-Farnesene	0.024	percent	
terpenes	Caryophyllene Oxide	0.0213	percent	
terpenes	b-Ocimene	0.0174	percent	
terpenes	Camphene	0.0166	percent	
terpenes	Borneol	0.0132	percent	
terpenes	Valencene	0.01	percent	
terpenes	Terpinolene	0.0061	percent	
terpenes	Sabinene Hydrate	0.0054	percent	
cannabinoids	THCa	26.875	percent	
cannabinoids	D9-THC	1.15	percent	
cannabinoids	CBGa	0.96	percent	
cannabinoids	CBCa	0.43	percent	
cannabinoids	THCVa	0.13	percent	
cannabinoids	CBDa	0.126	percent	
cannabinoids	CBG	<LOQ	percent	
cannabinoids	CBC	<LOQ	percent	
cannabinoids	D8-THC	ND	percent	
cannabinoids	THCV	ND	percent	
cannabinoids	CBD	ND	percent	
cannabinoids	CBDV	ND	percent	
pesticides	Cyfluthrin	ND	µg/g	PASS
pesticides	Cypermethrin	ND	µg/g	PASS
pesticides	Diazinon	ND	µg/g	PASS
pesticides	Dimethomorph	ND	µg/g	PASS
pesticides	Etoxazole	ND	µg/g	PASS
pesticides	Fenhexamid	ND	µg/g	PASS
pesticides	Fenpyroximate	ND	µg/g	PASS
pesticides	Flonicamid	ND	µg/g	PASS
pesticides	Fludioxonil	ND	µg/g	PASS
pesticides	Hexythiazox	ND	µg/g	PASS
pesticides	Imidacloprid	ND	µg/g	PASS
pesticides	Kresoxim-methyl	ND	µg/g	PASS

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Download - saved our CoA data to an .xlsx worksheet

coa-data-2022-08-26-17-50-48.xlsx



Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

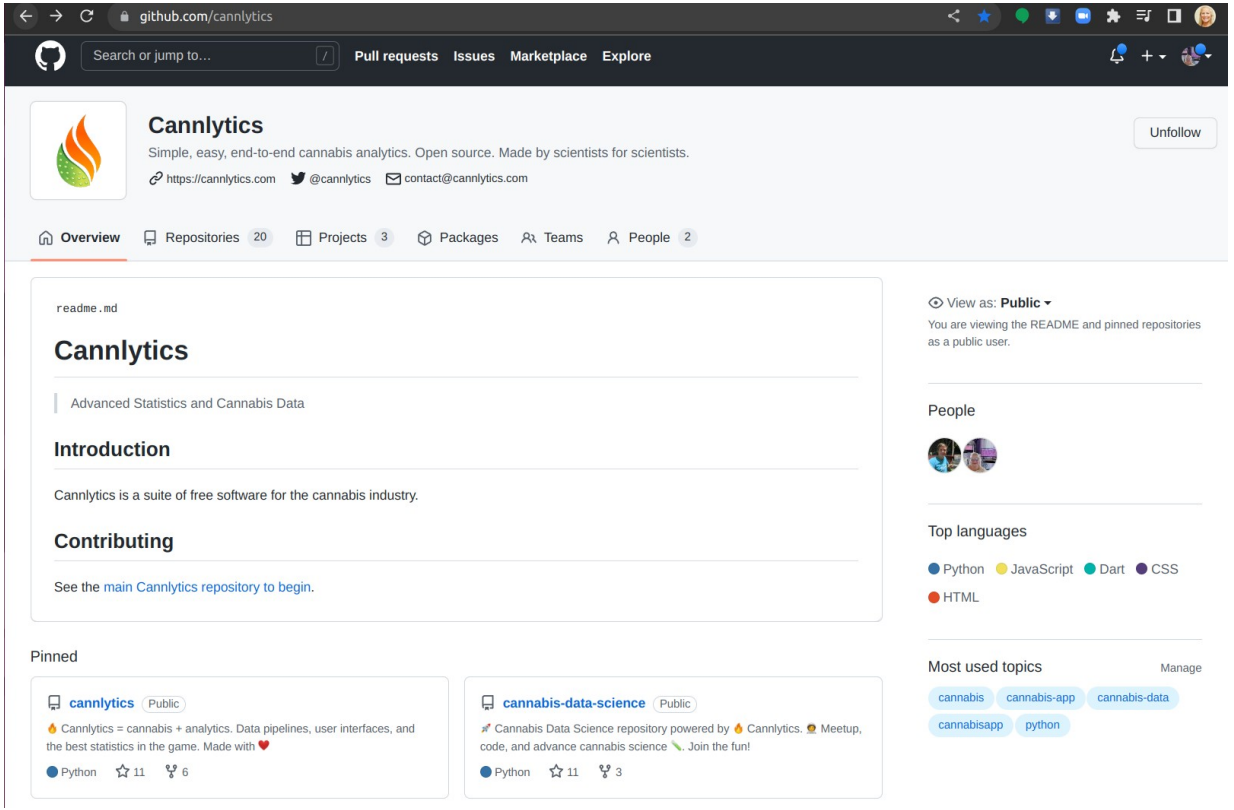
Let’s Start the Debugger and step thru some code…

https://github.com/cannlytics/

- web_scraping_script.py (Selenium demo)
- cs_labs.py (Beautiful Soup demo)
- coa.py (pdf_plumber demo)

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Code is available on https://github.com/cannlytics/



Cannlytics main repository

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

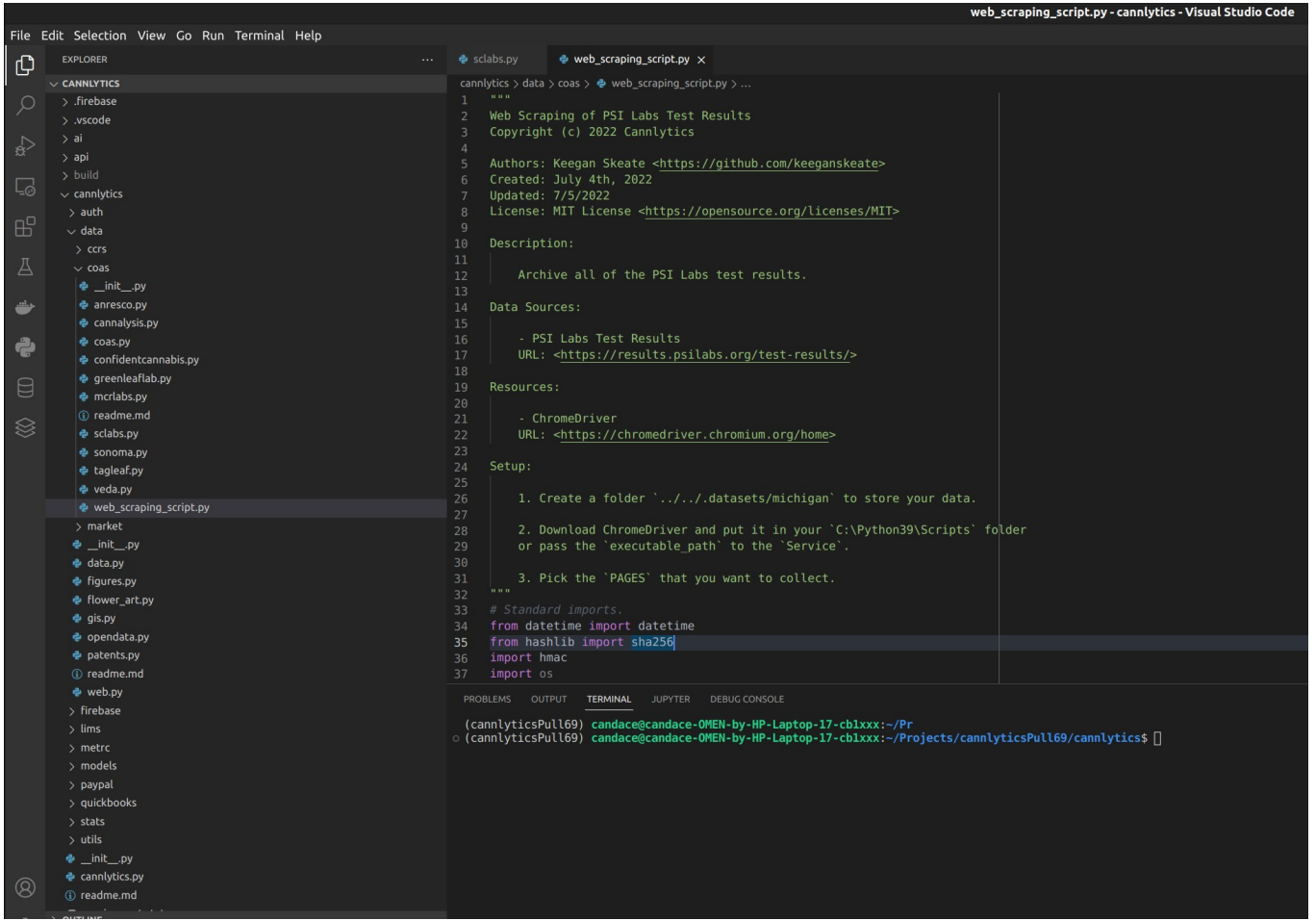
THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Please cite the following if you use the code examples in your research:

```
@misc{cannlytics2022,
  title={Cannabis Data Science},
  author={Skeate, Keegan and Rice, Charles and O'Sullivan-Sutherland, Candace},
  journal={https://github.com/cannlytics/cannabis-data-science},
  year={2022}
}
```

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Machine Learning algorithms require lots of Data for training accuracy

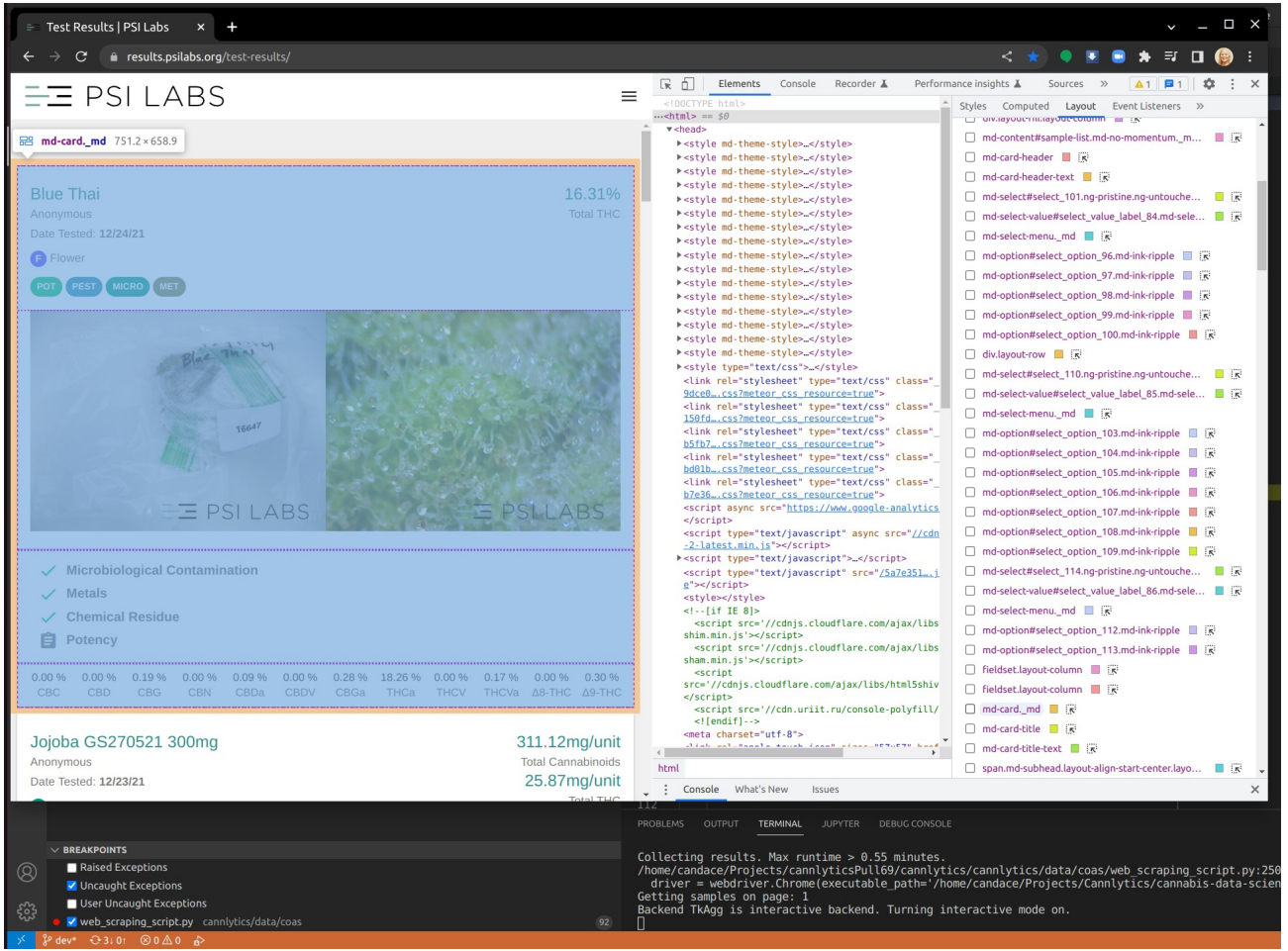


- Cannlytics collects and cleans available Cannabis USA data for robust datasets.
- There is much data to collect (Universities, Freedom of Information Act, Cannabis Lab Compliance Test facilities, NIH.gov… with 38 states legalizing medical Cannabis to-date...
- I will demo `web_scraping_script.py` to webscrape PSI Labs test data results using Selenium API with chrome driver for browser automation, working with core Javascript concepts(DOM) and handling AJAX/PJAX requests.
- PSI Labs operates in Cannabis legal USA states MI and CA.

  https://www.psilabs.org/

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Use Browser Inspector Tool to view website Elements



- Click on the three vertical dots on chromes top menu bar, choose More tools, then select Developer Tools

  OR

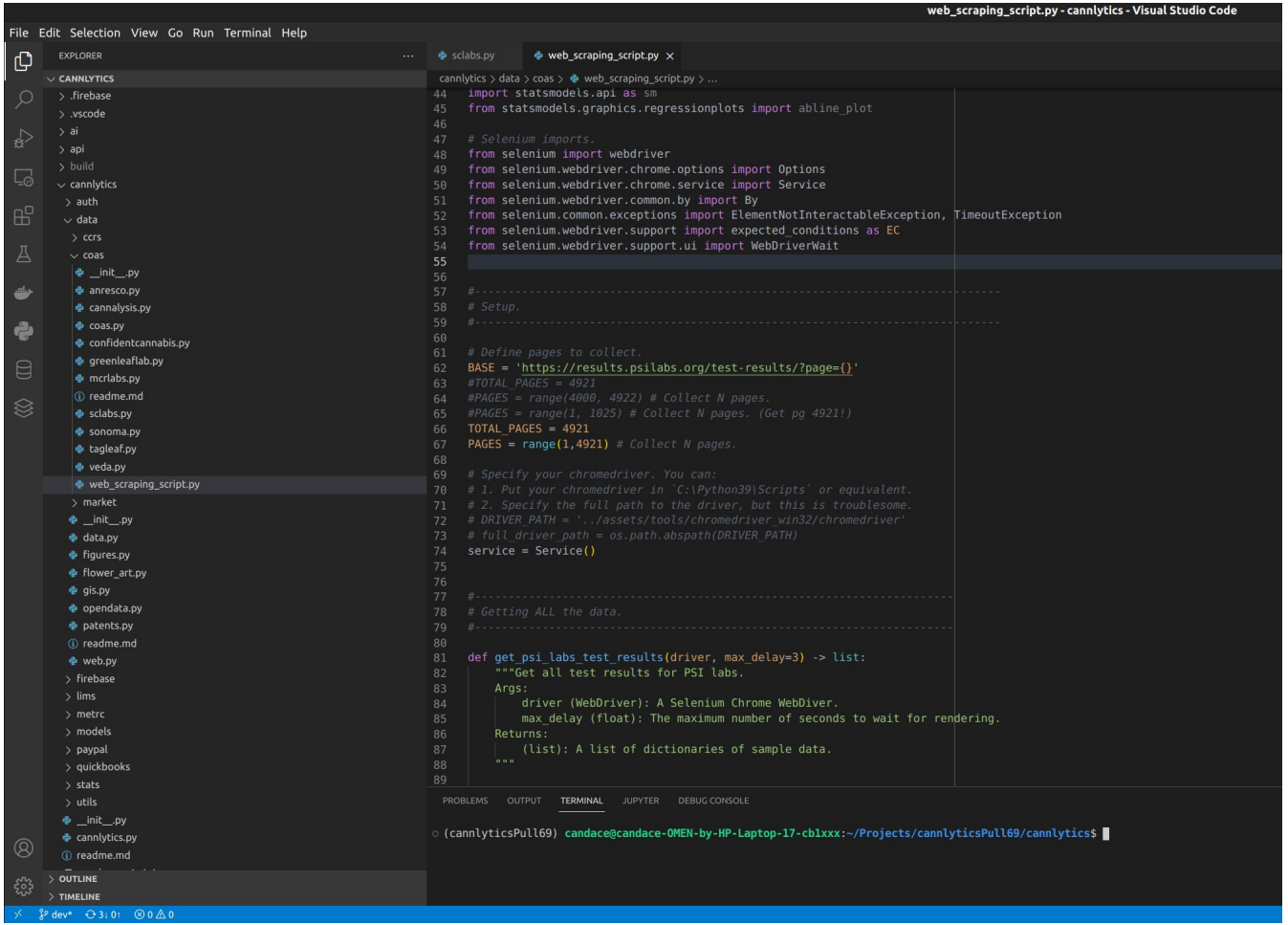
- Right-click on the web page and choose Inspect to access Developer Tools

  OR

- Use keyboard shortcuts Cntl-Shift+I or F12 for Windows or Linux and cmd+option+I for macOS users

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

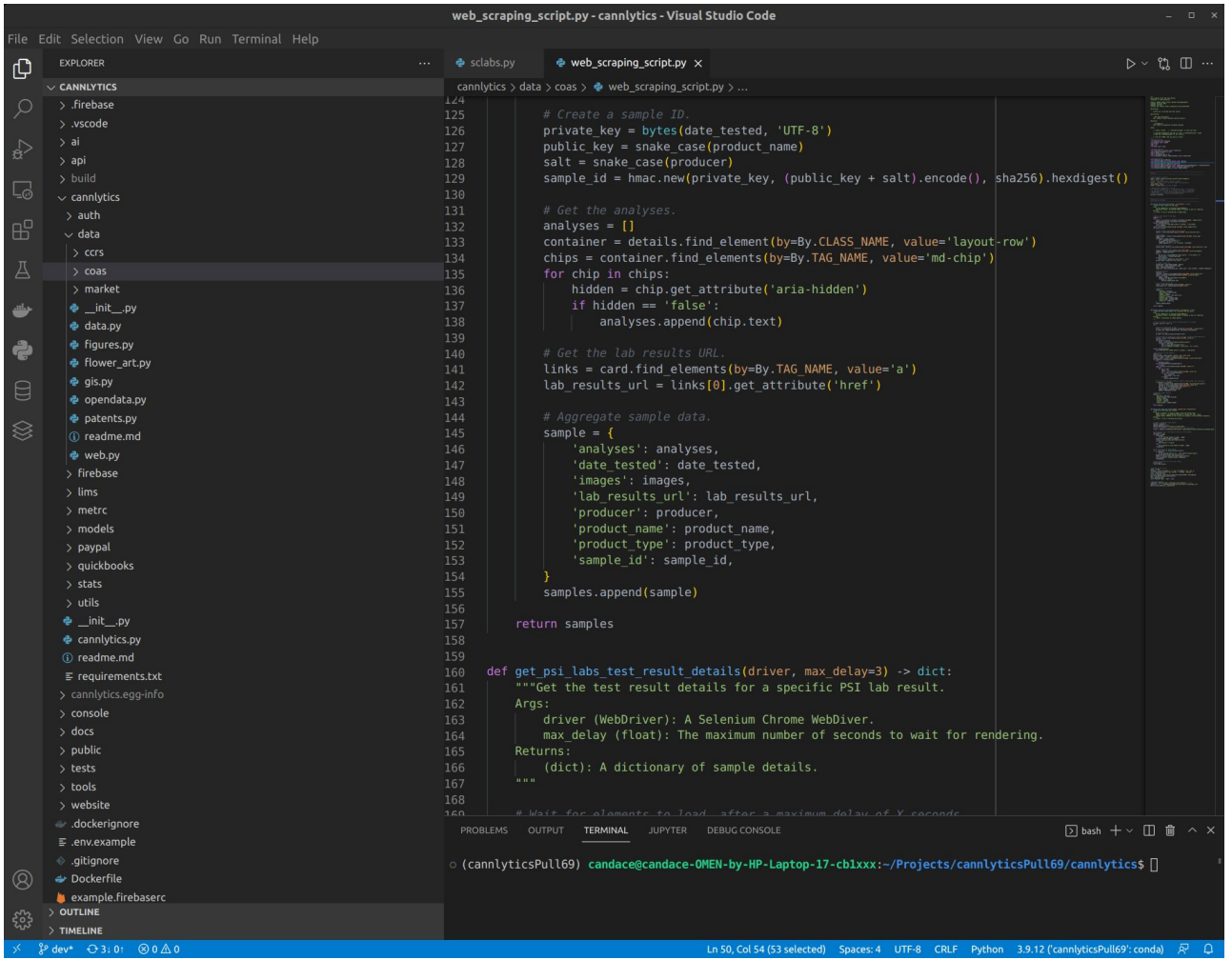
`web_scraping.py` Selenium Demo: specify chromedriver, declare variables, function `get_psi_labs_test_result` return List “samples”



- Specify chromedriver in c:\Python\Scripts or equivalent
  OR
  Specify the full path to the driver
  #DRIVER_PATH=`anaconda3/envs/web_scraping/bin/chromedriver` #full_driver_path= os.path.abspath(DRIVER_PATH)`
- Setup variables are declared
  Base URL, range of pages, total pages to collect
- Service = Service() is where we pass the Selenium Service object imported previously to use an instance of the Service() class from selenium.webdriver.chrome.service
- Define function `get_psi_labs_test_results` which collects test results for PSI labs and returns List “samples”
  - Get all the samples on the page
  - Get sample details from card
    - Images, product_name, producer, date tested, product_type
  - Create sample ID
  - Get the analyses
  - Get the Lab results URL
  - Aggregate “sample” data then append to/return List “samples” data

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

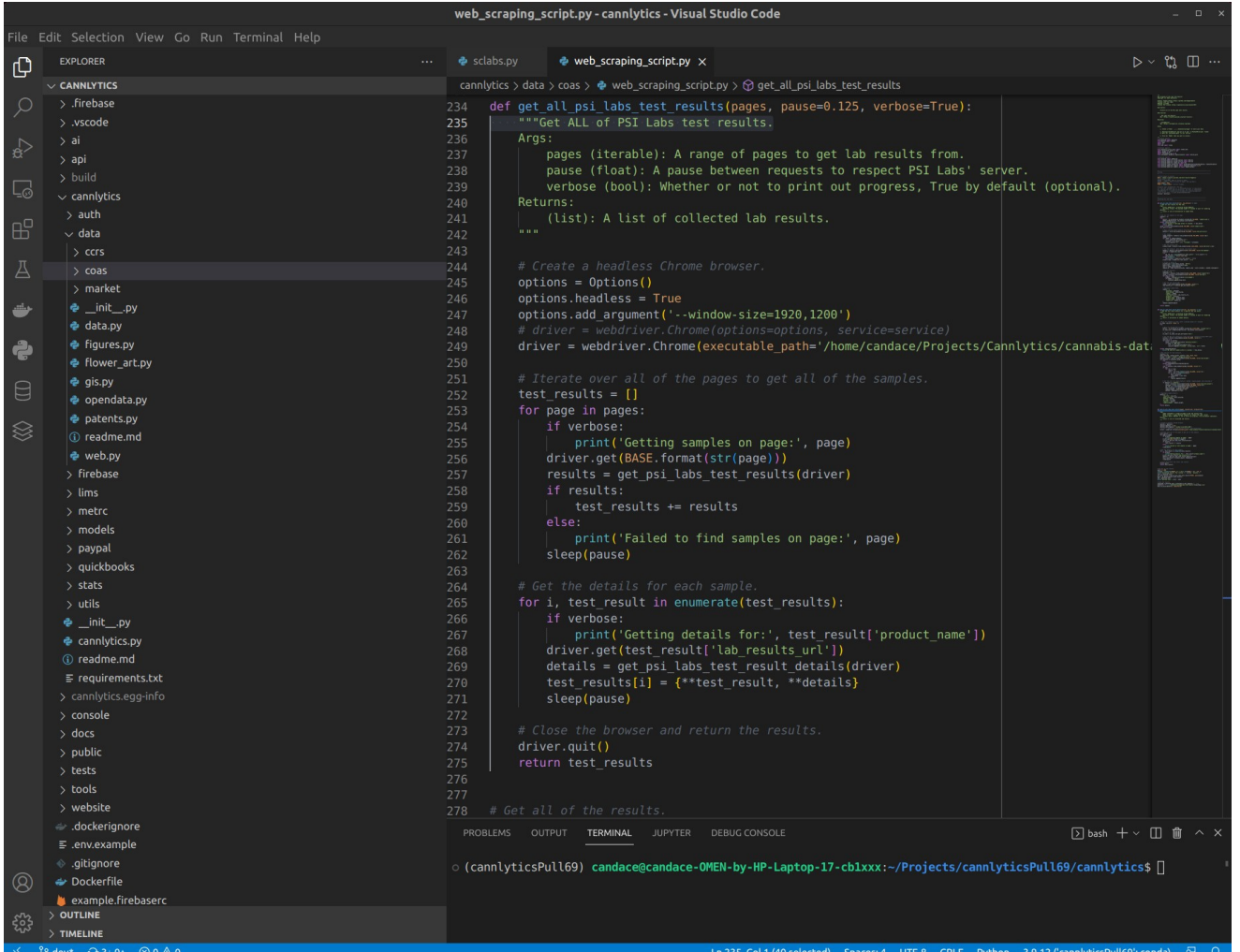
Declare function `get_psi_labs_test_details` return Dictionary “details”



- Define function `get_psi_labs_test_results_details` collects the test results details for a specific PSI labs test result and returns Dictionary “details”
- Wait for elements to load
- Wait for QR code to load then Get
- Get CoA Certificate of Analysis URLS
- Get Results for each analysis
- Aggregate sample details
- Return Dictionary “Details”

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

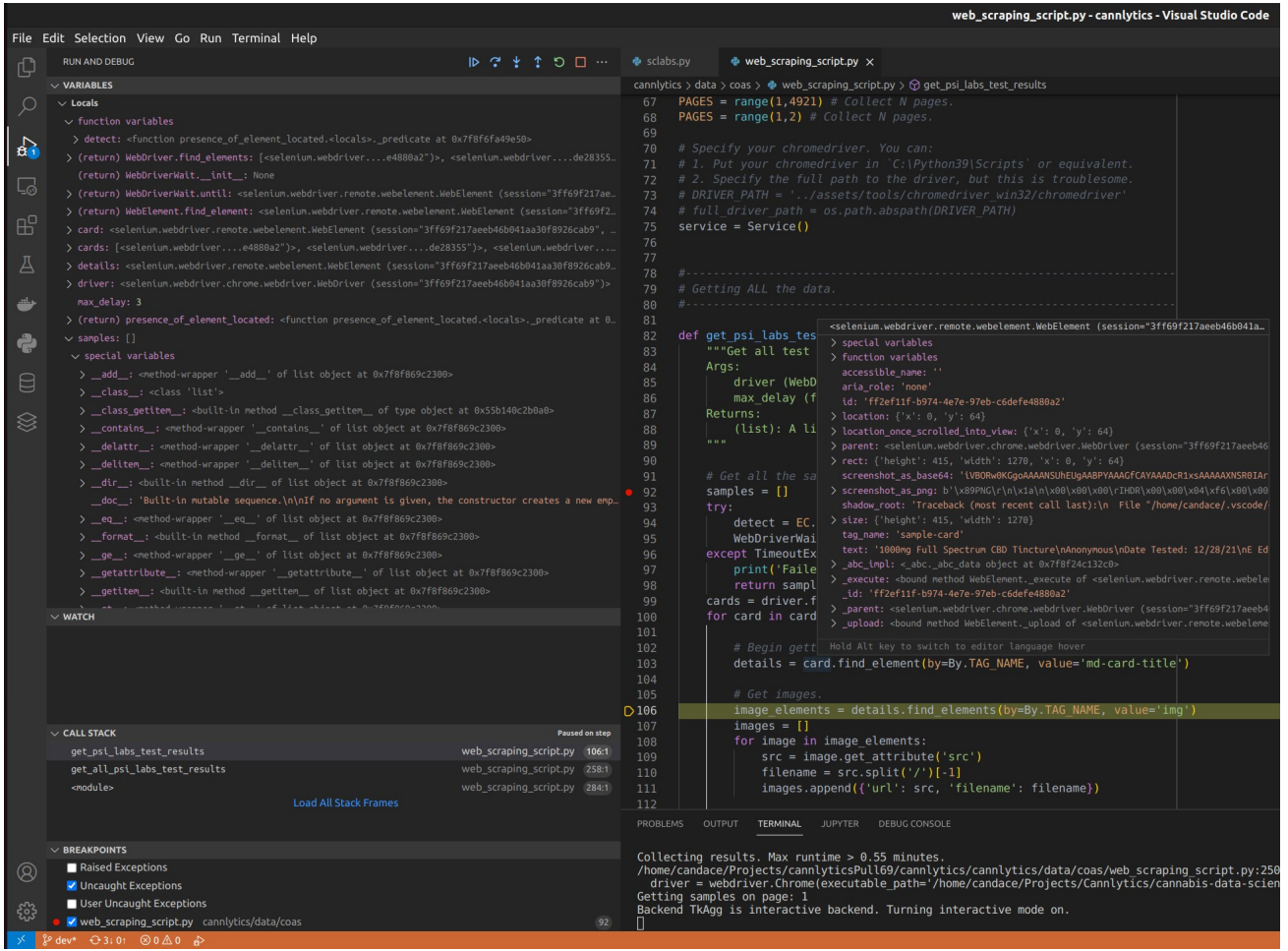
Declare function `get_all_psi_labs_test_result` return List “all_test_results”



- Define function `get_all_psi_labs_test_results` which collects ALL of PSI Labs’ test results and returns List “test_results”
- Create headless chrome browser
- Chromedriver gets the Base URL and Page
- Iterate over all the pages to place all the samples into Dictionary “test_results”
- Call `get_psi_labs_test_results` passing chromedriver to get “results” and add to “test_results”
- Call `get_psi_labs_test_result_details` passing chromedriver to get Dictionary “details” to add to List “test_results”
- “all_test_results” contains “test_results”
- Convert List “all_test_results” to Dataframe then write out to .xlsx file

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

VSCode Debugger used to view variable values



- Hover over a variable to see its value

  OR

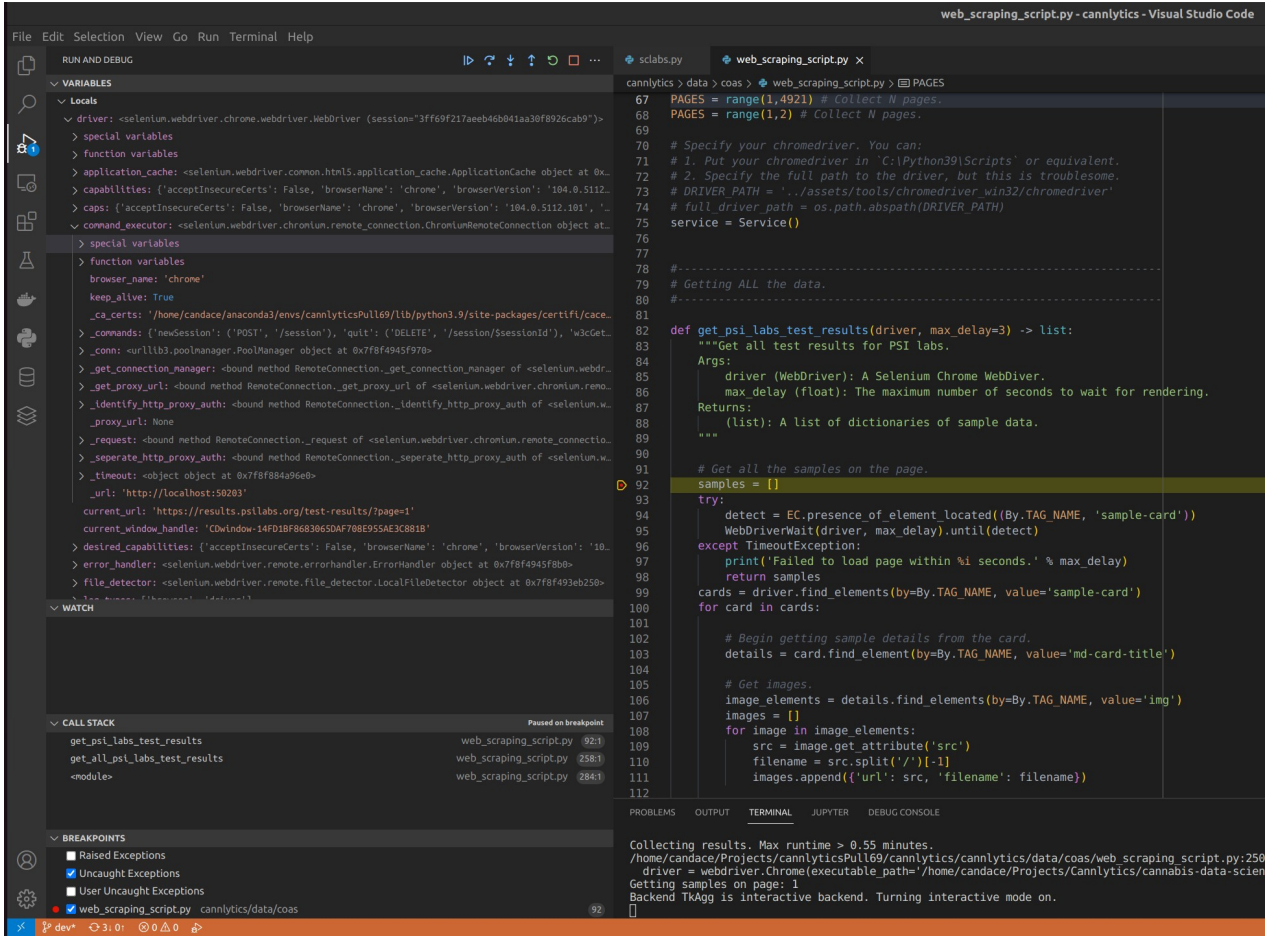
- Right-click on the web page and choose Inspect to access Developer Tools

  OR

- Use keyboard shortcuts Cntl-Shift+I or F12 for Windows or Linux and cmd+option+I for macOS users

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

VSCode Breakpoint(s) used to pause debugger execution



- Breakpoint set at the beginning of `get_psi_labs_test_results` line 92 samples = [] called from `get_all_psi_labs_test_results` called from <module>

- view “State of Code”
  - Variables
  - Locals
  - driver:
  - command_executor:
  - current_url: ‘https://results.psilabs.org/test-results/?page=1’

- view “Call Stack”
  - get_psi_labs_test_results
  - get_all_psi_labs_test_results
  - <module>

- Step into the code

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team

Let’s step thru the Selenium web scraping code…

… stay tuned for Beautiful Soup and pdfplumber code snippets and code step thru

Copyright (c) 2020-2022 Cannlytics and the Cannabis Data Science Team