2023

Exploratory Data Analysis in Jupyter Notebook inside ULEAD

ACCESSING THE TOOL AND UNDERSTANDING THE FUNCTIONS AND FEATURES

CENAN PIRANI & JAVIER SANZ

Document version 1.0

Exploratory Data Analysis Jupyter Notebook

This guide describes how to use a tool for the initial exploration of Electronic Health Record (EHR) data that was generated by the CTSI and delivered to investigators for research purposes. The data for your project has been exported as a set of text files (using the "CSV" format). If you are using the ULEAD enclave, these have been placed in the 'Data' folder for this project.

This toolkit has then been built to contain commands that will describe each of the specific files in your dataset . The software library is written in Python using the Pandas data analytics library on the Jupyter Notebook platform and if you are familiar with Pandas you can proceed to conduct additional analyses in the same notebook.

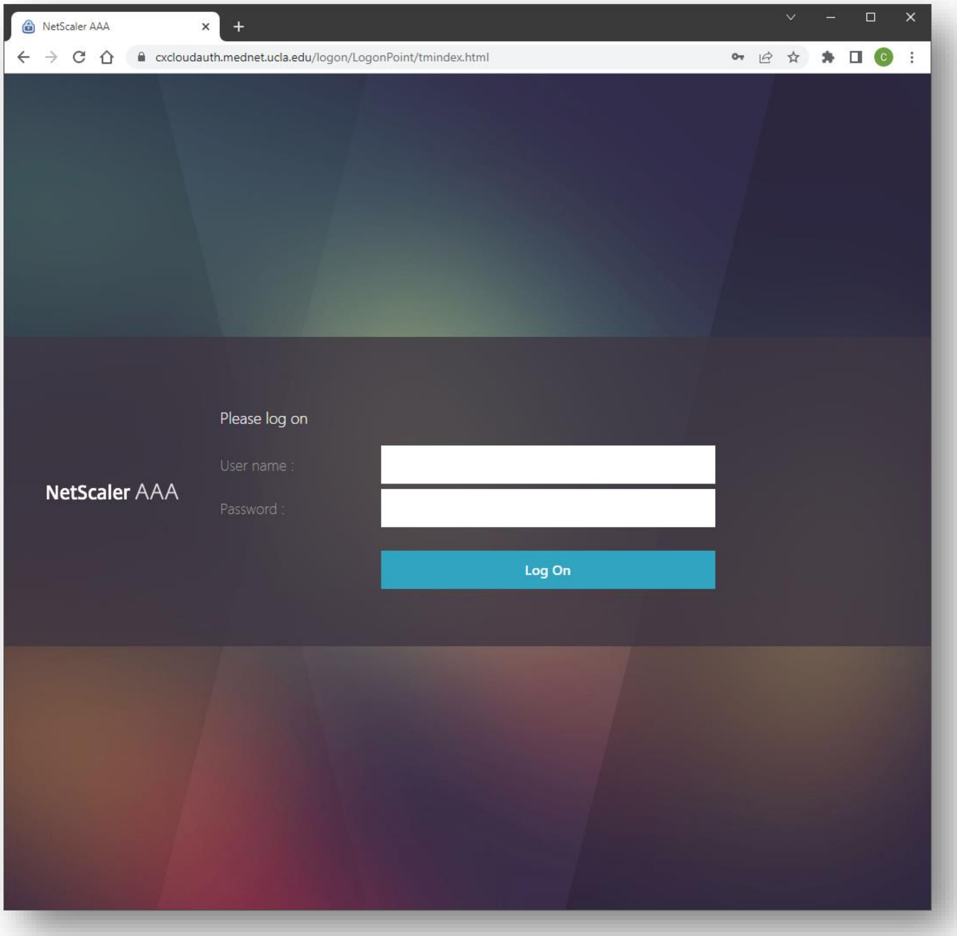
The tool is intended to help Investigators gain a deeper understanding of the data that they are receiving with tools that make it easy to complete basic and some intermediary data analysis tasks. It also aims to educate and encourage Investigators to expand their knowledge of analytics using Python and Pandas.

THE EDA NOTEBOOK FUNCTIONS

The notebook's main library, 'ehr_dp_lib.py', is made up of a set of functions developed by our team that targets the main descriptive statistics that investigators typically need to begin understanding their data. The code blocks in the notebook are pre-generated to the match the different files and variables contained in your specific data set. The functions are as follows but rather than reviewing these in detail here you can proceed to try examples using your own data:

- **describe_tables():** Returns a dataframe listing all the files in the 'Data' folder including row and column counts and descriptions
- **missingness(dataframe name):** Returns a dataframe of the number of null values per column.
- **catbar(dataframe name, column name, graph=(True or False))**: [Generated on categorical data type only] Returns a dataframe of counts of all the groups of categories in the specific column in the dataframe. When graph argument set to True returns a bar graph.
- **numstats(dataframe name, column name):** [Generated on number data type only] Returns a dataframe of descriptive statistics (ie. mean, max, min, median, quartiles) for the column data.
- **dateline(dataframe name, column name):** [Generated on date data type only] Returns a line graph of the freuency of specific dates along an x-axis of time.

- **flow_stats(flowsheet dataframe):** [Generated only if Flowsheet_Vitals.csv table in Data folder] Returns a dataframe of descriptive statistics for common vitals sign types (ie. Height, Weight, Temperature, Sp02, Pulse, BMI, Respirations).
- **lab_stats(lab dataframe, top=(10 or greater)):** [Generated only if Labs.csv table in Data folder] Returns a dataframe of descriptive statistics for top lab procedures in dataset. The top argument can be adjusted to capture more lab procedures.
- **text_search(dataframe name, column name, text to search, ignore case=(True by default can also be set to False)):** Returns a dataframe based on a free text search of a specific column in an existing dataframe.



ACCESSING ULEAD WITH ANACONDA AND JUPYTER NOTEBOOK

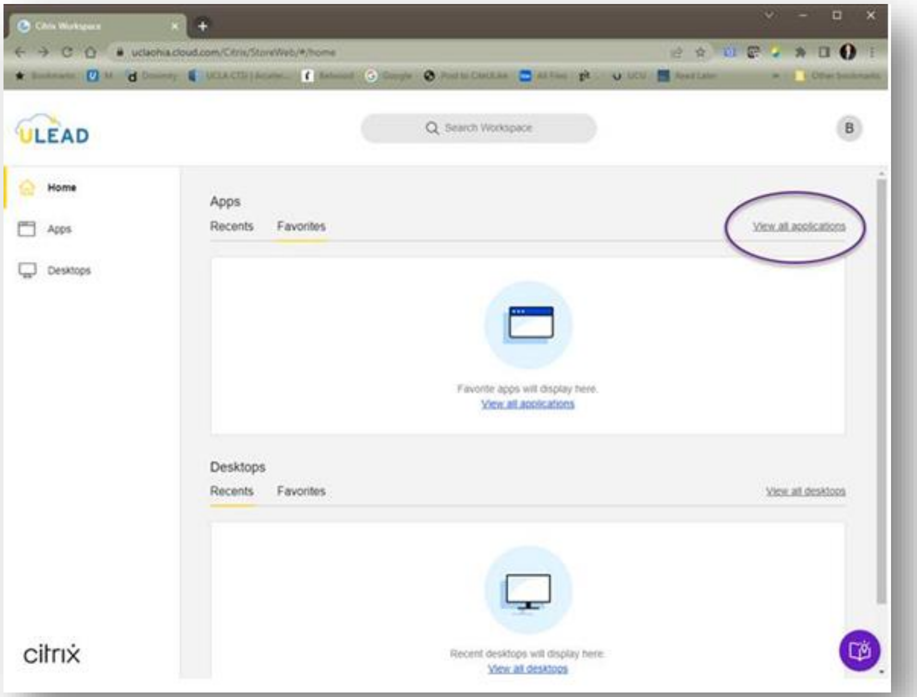
STEP 1: Launch ULEAD Citrix Environment

Navigate to: https://uclaohia.cloud.com/Citrix/StoreWeb/

The browser will re-rout and will present a login, enter your AD username / password. Logging in will also prompt the 2FA with Duo.

STEP 2: Open Anaconda Prompt

Once logged in, you will see the ULEAD Apps screen. If not already listed under “Recent” or “Favorites,” click on the link: ‘View all applications.’

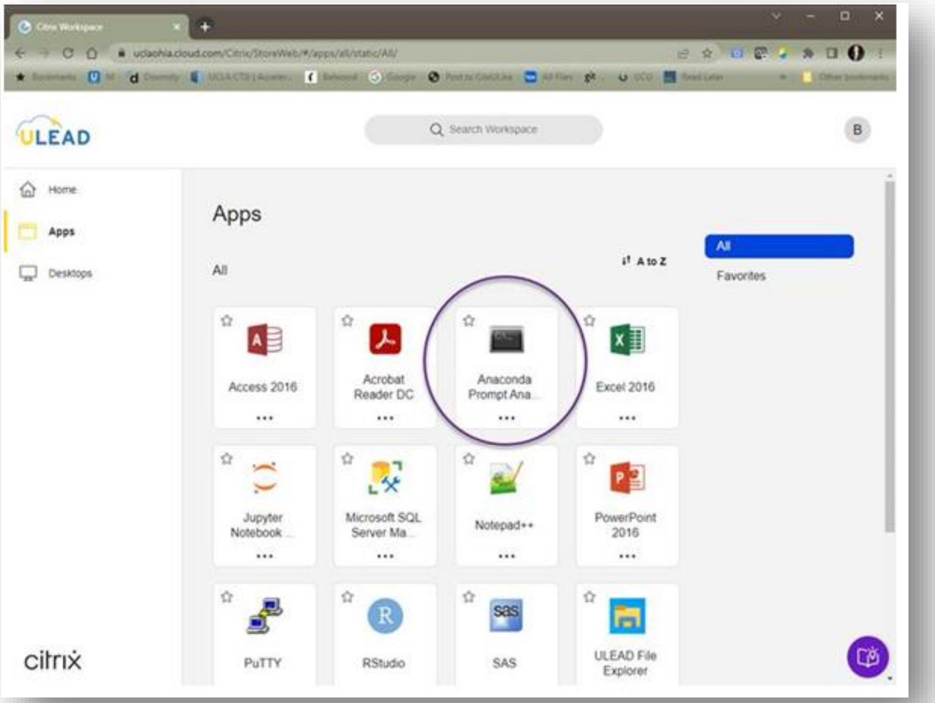


Favorite apps will display here.
View all applications

Recent desktops will display here.
View all desktops

Then click on the icon ‘Anaconda Prompt.’

(you might want to click on the ★ next to the icon and mark it as ‘favorite’ for future sessions).



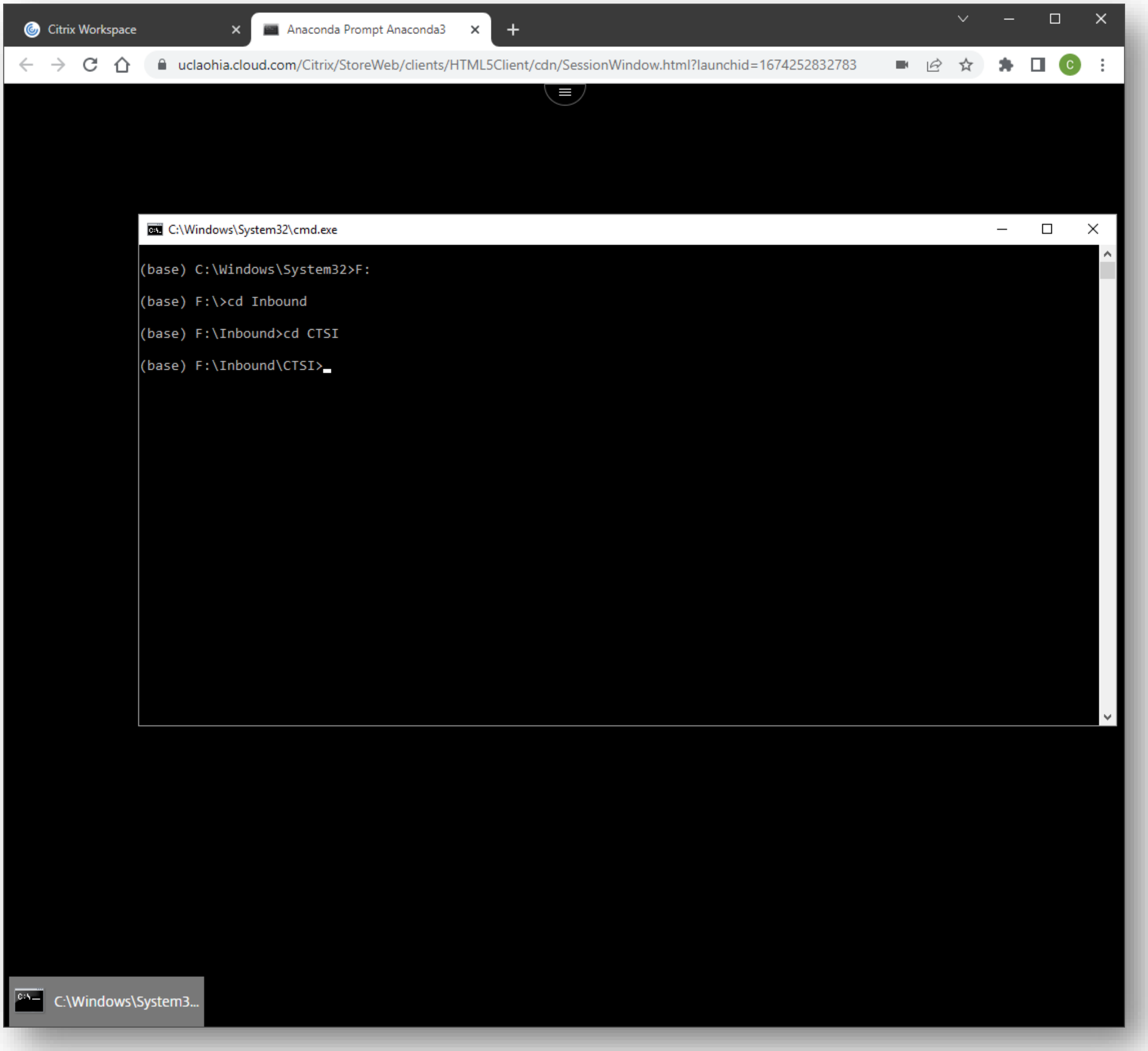
STEP 3: Navigate to root folder

Once finished loading a prompt will appear in the window in a virtual screen.

In the prompt enter the following commands:

- F: <enter>
- cd Inbound <enter>
- cd CTSI <enter>

At this point the prompt should look like this:


Citrix Workspace
Anaconda Prompt Anaconda3
uclaohia.cloud.com/Citrix/StoreWeb/clients/HTML5Client/cdn/SessionWindow.html?launchid=1674252832783
C:\Windows\System32\cmd.exe
C:\Windows\System3...

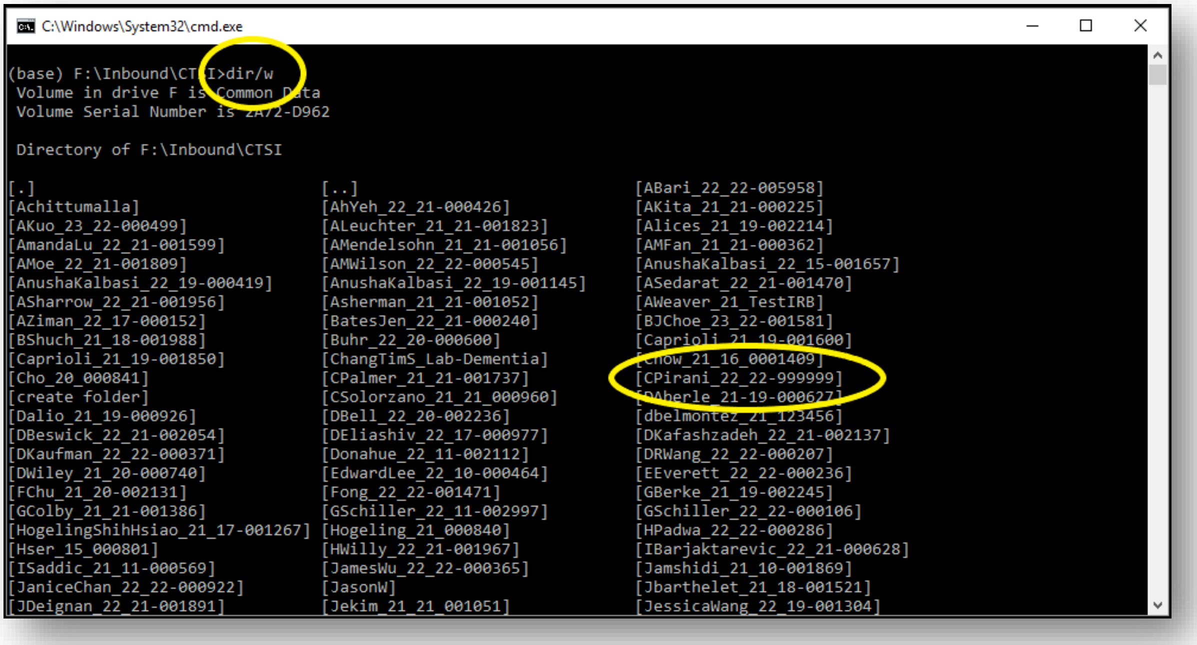
```
(base) C:\Windows\System32>F:

(base) F:\>cd Inbound

(base) F:\Inbound>cd CTSI

(base) F:\Inbound\CTSI>_
```

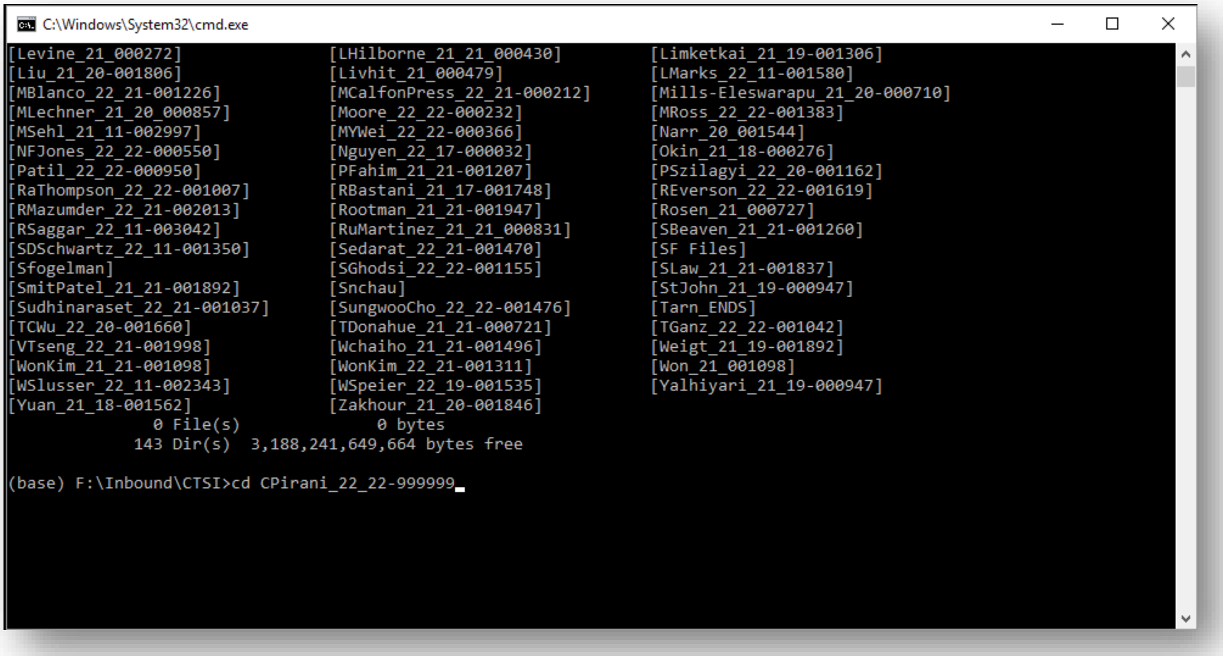
Now enter the command “dir/w” to list all of the folders in the directory, find your folder and enter: “cd <folder name> <enter>.”



```
(base) F:\Inbound\CTSI>dir/w
 Volume in drive F is Common Data
 Volume Serial Number is 2A72-D962

 Directory of F:\Inbound\CTSI

[.]                              [..]                             [ABari_22_22-005958]
[Achittumalla]                   [AhYeh_22_21-000426]             [AKita_21_21-000225]
[AKuo_23_22-000499]              [ALeuchter_21_21-001823]         [Alices_21_19-002214]
[AmandaLu_22_21-001599]          [AMendelsohn_21_21-001056]       [AMFan_21_21-000362]
[AMoe_22_21-001809]              [AMWilson_22_22-000545]          [AnushaKalbasi_22_15-001657]
[AnushaKalbasi_22_19-000419]     [AnushaKalbasi_22_19-001145]     [ASedarat_22_21-001470]
[ASharrow_22_21-001956]          [Asherman_21_21-001052]          [AWeaver_21_TestIRB]
[AZiman_22_17-000152]            [BatesJen_22_21-000240]          [BJChoe_23_22-001581]
[BShuch_21_18-001988]            [Buhr_22_20-000600]              [Caprioli_21_19-001600]
[Caprioli_21_19-001850]          [ChangTimS_Lab-Dementia]         [Chow_21_16_0001409]
[Cho_20_000841]                  [CPalmer_21_21-001737]           [CPirani_22_22-999999]
[create folder]                  [CSolorzano_21_21_000960]        [DAberle_21-19-000627]
[Dalio_21_19-000926]             [DBell_22_20-002236]             [dbelmontez_21_123456]
[DBeswick_22_21-002054]          [DEliashiv_22_17-000977]         [DKafashzadeh_22_21-002137]
[DKaufman_22_22-000371]          [Donahue_22_11-002112]           [DRWang_22_22-000207]
[DWiley_21_20-000740]            [EdwardLee_22_10-000464]         [EEverett_22_22-000236]
[FChu_21_20-002131]              [Fong_22_22-001471]              [GBerke_21_19-002245]
[GColby_21_21-001386]            [GSchiller_22_11-002997]         [GSchiller_22_22-000106]
[HogelingShihHsiao_21_17-001267] [Hogeling_21_000840]             [HPadwa_22_22-000286]
[Hser_15_000801]                 [HWilly_22_21-001967]            [IBarjaktarevic_22_21-000628]
[ISaddic_21_11-000569]           [JamesWu_22_22-000365]           [Jamshidi_21_10-001869]
[JaniceChan_22_22-000922]        [JasonW]                         [Jbarthelet_21_18-001521]
[JDeignan_22_21-001891]          [Jekim_21_21_001051]             [JessicaWang_22_19-001304]
```

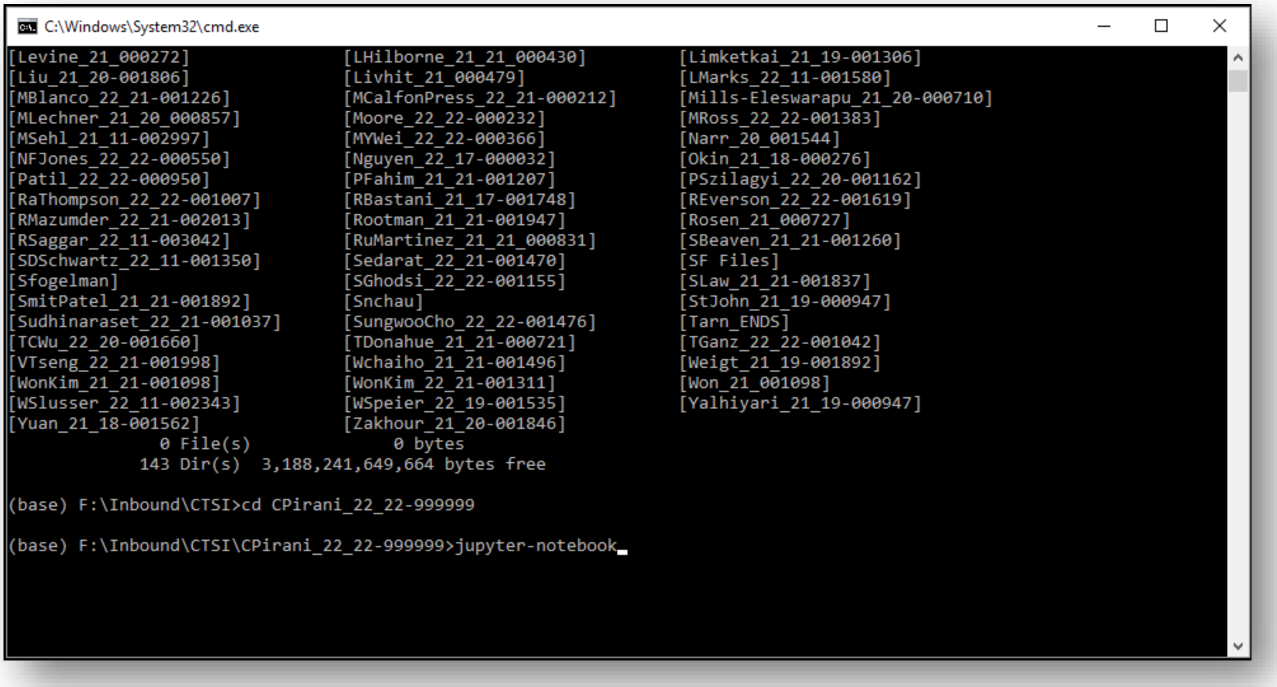


```
[Levine_21_000272]               [LHilborne_21_21_000430]         [Limketkai_21_19-001306]
[Liu_21_20-001806]               [Livhit_21_000479]               [LMarks_22_11-001580]
[MBlanco_22_21-001226]           [MCalfonPress_22_21-000212]      [Mills-Eleswarapu_21_20-000710]
[MLechner_21_20_000857]          [Moore_22_22-000232]             [MRoss_22_22-001383]
[MSehl_21_11-002997]             [MYWei_22_22-000366]             [Narr_20_001544]
[NFJones_22_22-000550]           [Nguyen_22_17-000032]            [Okin_21_18-000276]
[Patil_22_22-000950]             [PFahim_21_21-001207]            [PSzilagyi_22_20-001162]
[RaThompson_22_22-001007]        [RBastani_21_17-001748]          [REverson_22_22-001619]
[RMazumder_22_21-002013]         [Rootman_21_21-001947]           [Rosen_21_000727]
[RSaggar_22_11-003042]           [RuMartinez_21_21_000831]        [SBeaven_21_21-001260]
[SDSchwartz_22_11-001350]        [Sedarat_22_21-001470]           [SF Files]
[Sfogelman]                      [SGhodsi_22_22-001155]           [SLaw_21_21-001837]
[SmitPatel_21_21-001892]         [Snchau]                         [StJohn_21_19-000947]
[Sudhinaraset_22_21-001037]      [SungwooCho_22_22-001476]        [Tarn_ENDS]
[TCWu_22_20-001660]              [TDonahue_21_21-000721]          [TGanz_22_22-001042]
[VTseng_22_21-001998]            [Wchaiho_21_21-001496]           [Weigt_21_19-001892]
[WonKim_21_21-001098]            [WonKim_22_21-001311]            [Won_21_001098]
[WSlusser_22_11-002343]          [WSpeier_22_19-001535]           [Yalhiyari_21_19-000947]
[Yuan_21_18-001562]              [Zakhour_21_20-001846]
               0 File(s)              0 bytes
             143 Dir(s)  3,188,241,649,664 bytes free

(base) F:\Inbound\CTSI>cd CPirani_22_22-999999_
```

STEP 5: Launch Jupyter Notebook application

Now type the command “jupyter-notebook” and press <enter>. After a short time it will open a Chrome browser with the Jupyter Notebook explorer in your own ULEAD folder.



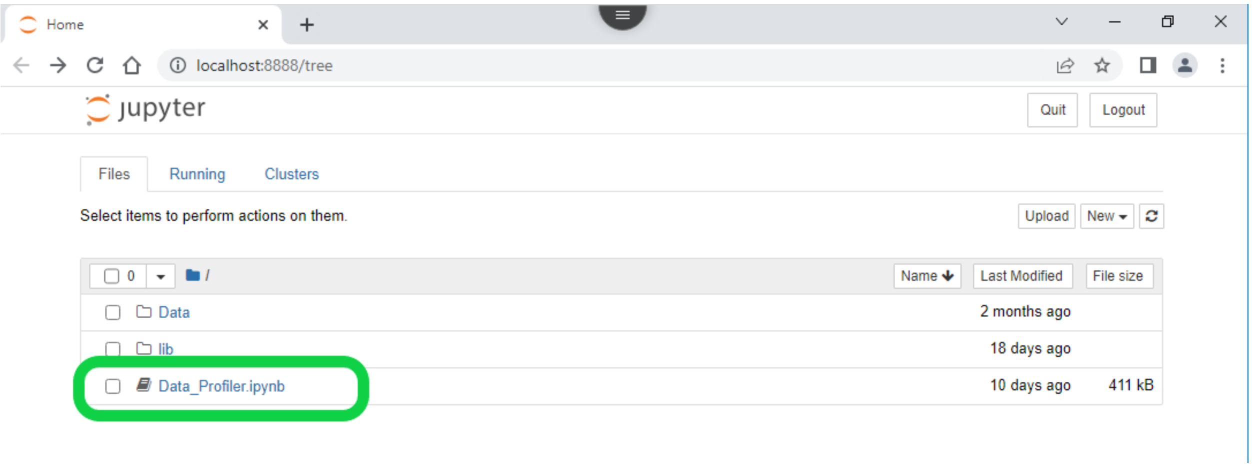
```
[Levine_21_000272]               [LHilborne_21_21_000430]         [Limketkai_21_19-001306]
[Liu_21_20-001806]               [Livhit_21_000479]               [LMarks_22_11-001580]
[MBlanco_22_21-001226]           [MCalfonPress_22_21-000212]      [Mills-Eleswarapu_21_20-000710]
[MLechner_21_20_000857]          [Moore_22_22-000232]             [MRoss_22_22-001383]
[MSehl_21_11-002997]             [MYWei_22_22-000366]             [Narr_20_001544]
[NFJones_22_22-000550]           [Nguyen_22_17-000032]            [Okin_21_18-000276]
[Patil_22_22-000950]             [PFahim_21_21-001207]            [PSzilagyi_22_20-001162]
[RaThompson_22_22-001007]        [RBastani_21_17-001748]          [REverson_22_22-001619]
[RMazumder_22_21-002013]         [Rootman_21_21-001947]           [Rosen_21_000727]
[RSaggar_22_11-003042]           [RuMartinez_21_21_000831]        [SBeaven_21_21-001260]
[SDSchwartz_22_11-001350]        [Sedarat_22_21-001470]           [SF Files]
[Sfogelman]                      [SGhodsi_22_22-001155]           [SLaw_21_21-001837]
[SmitPatel_21_21-001892]         [Snchau]                         [StJohn_21_19-000947]
[Sudhinaraset_22_21-001037]      [SungwooCho_22_22-001476]        [Tarn_ENDS]
[TCWu_22_20-001660]              [TDonahue_21_21-000721]          [TGanz_22_22-001042]
[VTseng_22_21-001998]            [Wchaiho_21_21-001496]           [Weigt_21_19-001892]
[WonKim_21_21-001098]            [WonKim_22_21-001311]            [Won_21_001098]
[WSlusser_22_11-002343]          [WSpeier_22_19-001535]           [Yalhiyari_21_19-000947]
[Yuan_21_18-001562]              [Zakhour_21_20-001846]
               0 File(s)              0 bytes
             143 Dir(s)  3,188,241,649,664 bytes free

(base) F:\Inbound\CTSI>cd CPirani_22_22-999999

(base) F:\Inbound\CTSI\CPirani_22_22-999999>jupyter-notebook_
```

STEP 6: Open EDA document

At this point you can click on the file: ‘Data_Profiler.ipynb’ to launch the notebook:



Select items to perform actions on them.

Name	Last Modified	File size
□ Data	2 months ago	
□ lib	18 days ago	
□ Data_Profiler.ipynb	10 days ago	411 kB