DSC 180B

2023/01/20 - Progress Update

Start processing the data collected

- **Python + SQL**
    - Read data from the databases
    - Process data
        - Combine datasets from
            - COUNTERS_STRING_TIME_DATA
            - COUNTERS_ULL_TIME_DATA
        - Conduct EDA
        - Visualize the data
    - Try to generalize the code as much as possible so that we can reuse it later
- **Tableau**
    - Visualize data

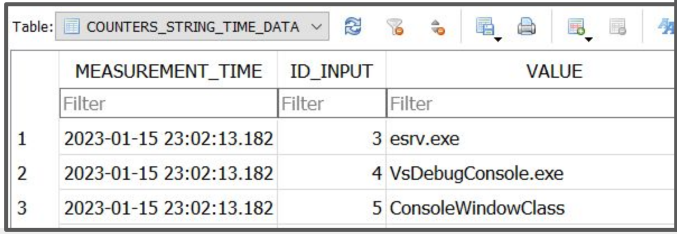


Table: COUNTERS_STRING_TIME_DATA

	MEASUREMENT_TIME	ID_INPUT	VALUE
	Filter	Filter	Filter
1	2023-01-15 23:02:13.182	3	esrv.exe
2	2023-01-15 23:02:13.182	4	VsDebugConsole.exe
3	2023-01-15 23:02:13.182	5	ConsoleWindowClass

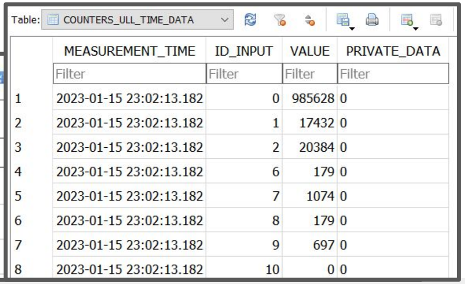


Table: COUNTERS_ULL_TIME_DATA

	MEASUREMENT_TIME	ID_INPUT	VALUE	PRIVATE_DATA
	Filter	Filter	Filter	Filter
1	2023-01-15 23:02:13.182	0	985628	0
2	2023-01-15 23:02:13.182	1	17432	0
3	2023-01-15 23:02:13.182	2	20384	0
4	2023-01-15 23:02:13.182	6	179	0
5	2023-01-15 23:02:13.182	7	1074	0
6	2023-01-15 23:02:13.182	8	179	0
7	2023-01-15 23:02:13.182	9	697	0
8	2023-01-15 23:02:13.182	10	0	0

Identify Data Quality Issues

- Identify why there is **“Missing Strings”**
  - Reasons: Realized it happens only when the classes are
    - “Shell_TrayWnd”
    - “ApplicationManager_DesktopShellWindow”,
    - “TaskListThumbnailWnd”
  - Is in relation to executable files
    - explorer.exe (no title when 1st opened)
    - chrome.exe
  - Solns:
    - Impute the data with the most used window name for that particular image

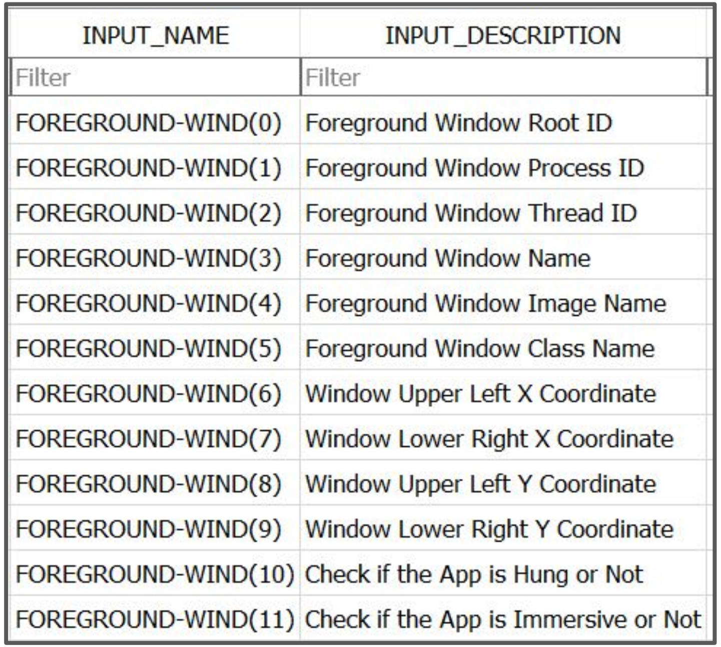
- Identify why there is “**Unable to Open Process**” in image names
  - Reasons: Realized this shows up for application which are run as administrator. (Ex: Command Prompt)
  - Solns: Can either drop or impute the data with cmd.exe

Identify Data Quality Issues

- **Too few entries** due to not actively switching between apps
    - Reasons: Often stay on Chrome to watch lectures or do HWs
    - Solns:
        - ● Can collect data when switching tabs
            - ○ Pros: relatively give more data
            - ○ Cons: when deployed in the field, might not want to collect the entire string of the tab b/c they can contain PIIs
        - ● Can do more data collection on different user computers/desktops.
    - Sanity check:
        - Regression of the newly collected data

Perform EDA

- Data collected



INPUT_NAME	INPUT_DESCRIPTION
Filter	Filter
FOREGROUND-WIND(0)	Foreground Window Root ID
FOREGROUND-WIND(1)	Foreground Window Process ID
FOREGROUND-WIND(2)	Foreground Window Thread ID
FOREGROUND-WIND(3)	Foreground Window Name
FOREGROUND-WIND(4)	Foreground Window Image Name
FOREGROUND-WIND(5)	Foreground Window Class Name
FOREGROUND-WIND(6)	Window Upper Left X Coordinate
FOREGROUND-WIND(7)	Window Lower Right X Coordinate
FOREGROUND-WIND(8)	Window Upper Left Y Coordinate
FOREGROUND-WIND(9)	Window Lower Right Y Coordinate
FOREGROUND-WIND(10)	Check if the App is Hung or Not
FOREGROUND-WIND(11)	Check if the App is Immersive or Not

- 2 data types
  - ULL
  - String

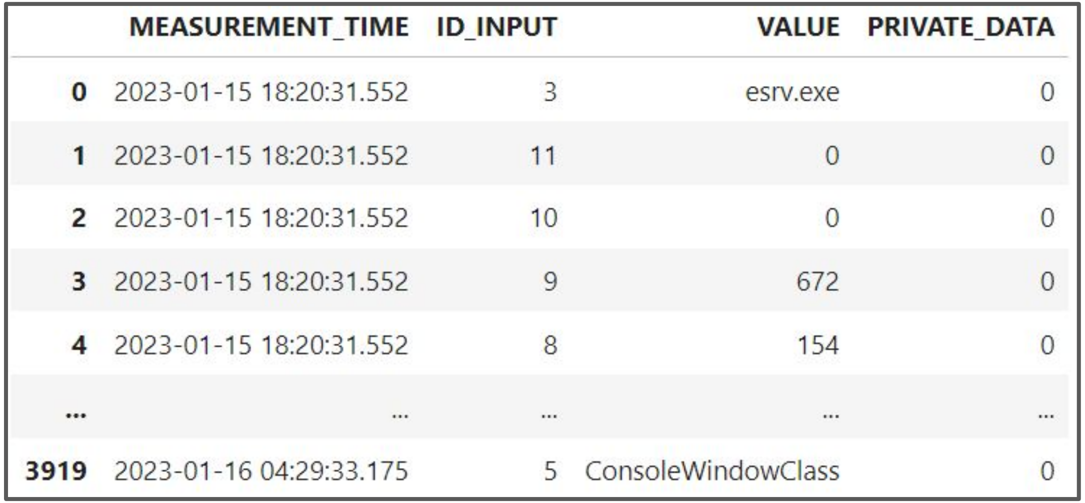
- 12 inputs related to the foreground window
  - Root/Process/Thread IDs,
  - Window/Image/Class Names,
  - Window Dimensions (rectangles)
  - App is Hung/Immersive or not

Perform EDA

- Python Dataframe:
  - 4 columns

MEASUREMENT_TIME	datetime64[ns]
ID_INPUT	int64
VALUE	object
PRIVATE_DATA	int32

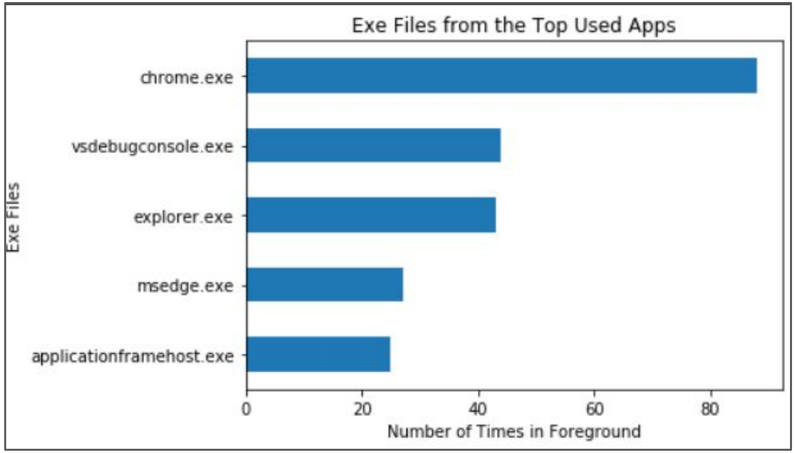
  - > 5000 rows in total



	MEASUREMENT_TIME	ID_INPUT	VALUE	PRIVATE_DATA
0	2023-01-15 18:20:31.552	3	esrv.exe	0
1	2023-01-15 18:20:31.552	11	0	0
2	2023-01-15 18:20:31.552	10	0	0
3	2023-01-15 18:20:31.552	9	672	0
4	2023-01-15 18:20:31.552	8	154	0
...	...	...	...	...
3919	2023-01-16 04:29:33.175	5	ConsoleWindowClass	0

Perform EDA

- As of Jan 19,
  - The number of **unique*** entries from both users is 1300+ and still counting
- Perform EDA on a subset of all databases
  - User 1, timeframe: Jan 15 - Jan 16
  - # Unique entries = 1308
  - 14 different types of exe files: ‘`vsdebugconsole.exe', 'explorer.exe', 'chrome.exe',` etc.



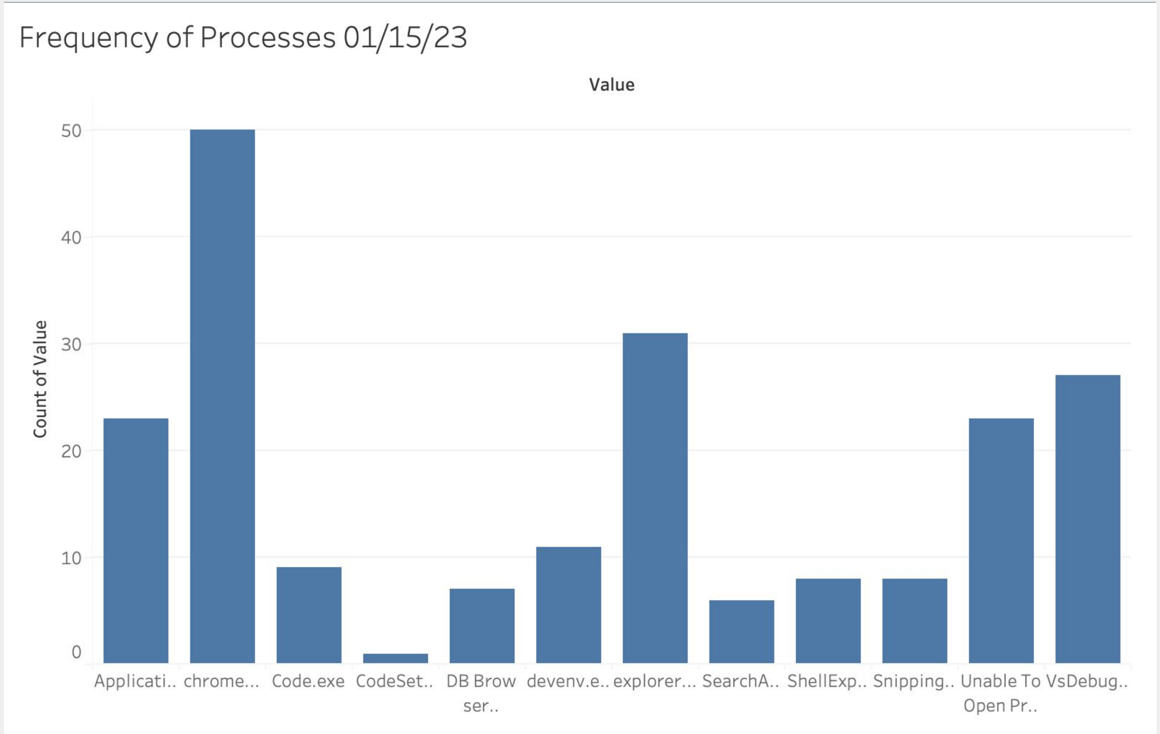
*unique: new data is recorded if a new tab/app is shown

Perform EDA

- Most Used Apps/Tabs
  - Chrome
    - Google Doc
    - Google Slides
    - Search
    - Mail
    - …
- PRIVATE_DATA: all 0s
- Average time used the device
  - User 1: 8 hrs (9 am - 1 pm, 8 pm - 12 am)
  - User 2: 8 hrs (2 pm - 10 pm)
- Time recorded was in UTC
- All apps are not hung
- Most apps are not immersive

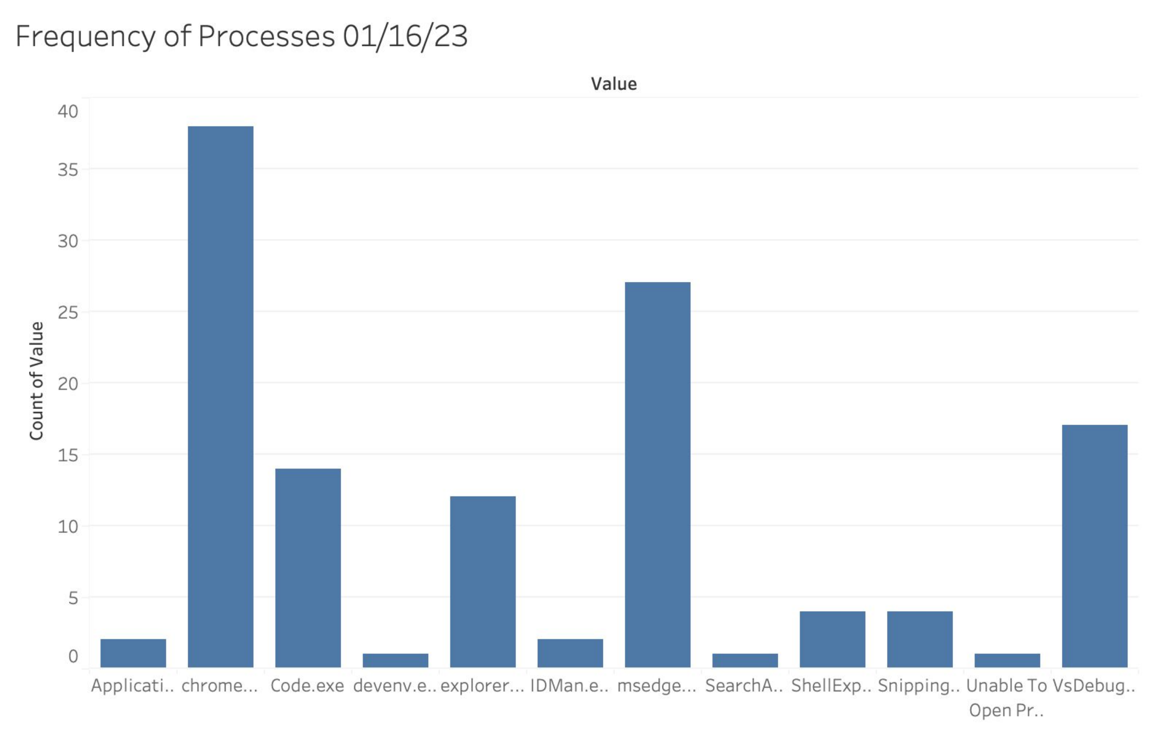
User 1, timeframe: Jan 15 - Jan 16

Visualization



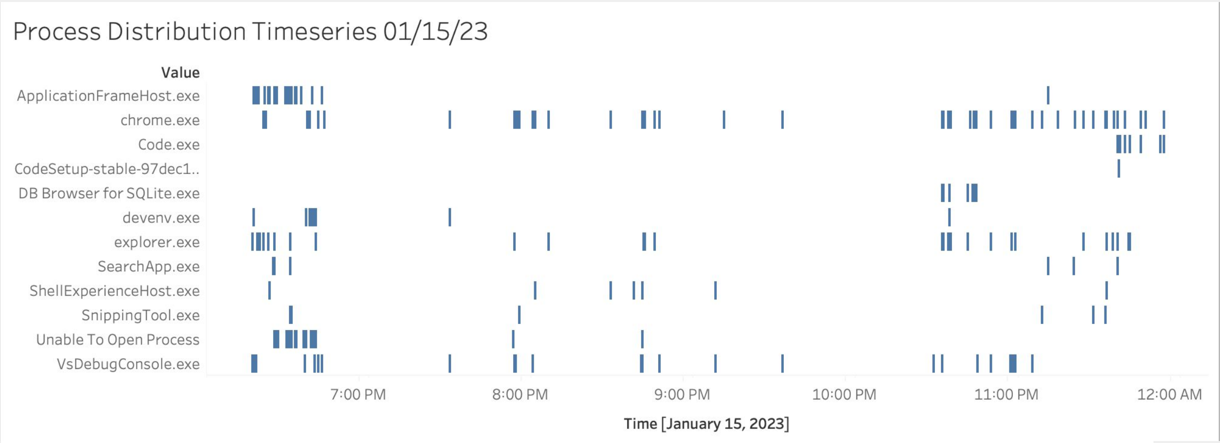
- Top 3 Frequency of Processes: chrome.exe > explorer.exe > VsDebugConsole.exe

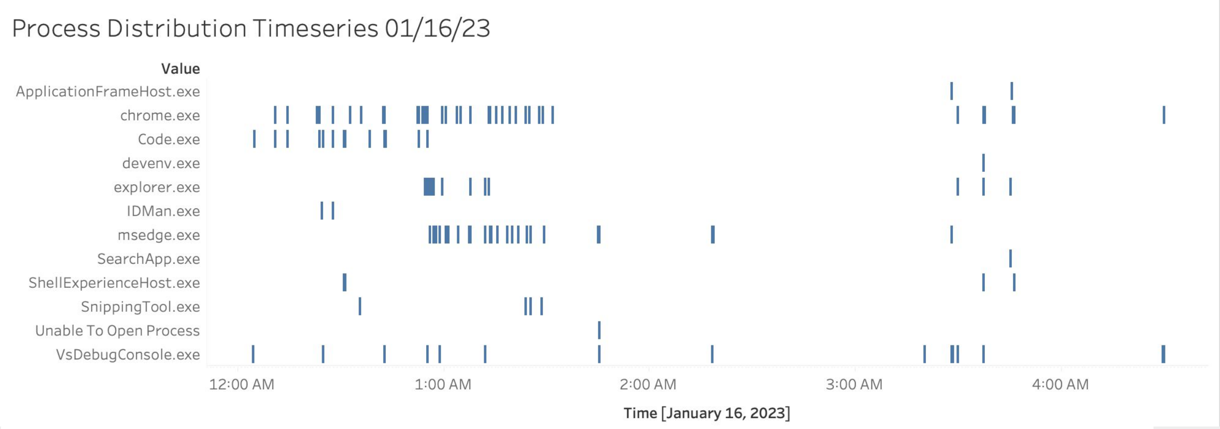
Visualization



- Top 3 Frequency of Processes: chrome.exe > msedge.exe > VsDebugConsole.exe
- Quite similar to Jan 15 data

Visualization





Source Code for EDA

https://github.com/miloncl/Dsc180b