# **Assignment**

# **Take Home Project - Data Scientist**

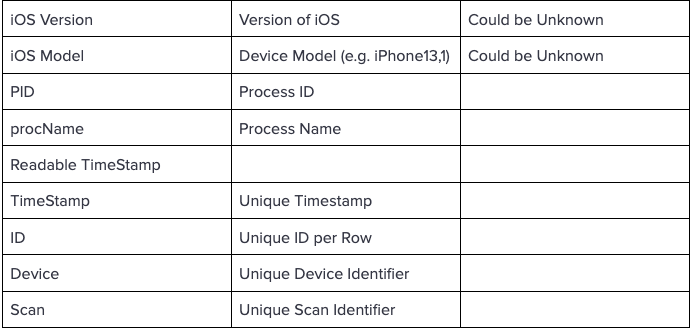
Expected Time of Analysis: No more than 3 hours  
Expected type of Output: Report & Visualisation  
Delivery: Upload to Ashby

Demonstrated Skills: Visualization of Data, Anomaly Detection  
Passing requirements: Submitted Report, Visual of Data, Detected Anomalies  
Evaluation Criteria: Checked files. Quality of the report.

Context

You are working as a Data Scientist in the Research Team for a security company called “Never Hacked Inc”. Your company recently got a huge sample of Process Data as a CSV Files.  
  
 Process Data here is a list of all the active Processes that were running on a device at a given timestamp. Each Process has a Process Name and a ProcessID (PID). Our assumption is that Malware will always run as a dedicated Process with its own Name. In the past Malware tried to hide themselves behind Process Names which look similar to existing ones,, but there are slight deviations. One example: e.g. rlaccountd (malicious) - roleaccountd (benign). But this might be over now. PIDs are assigned in a linear growing fashion for new processes. When a threshold is hit, it starts again at 1, but skips over running processes. PIDs should per process list.  
  
 The dataset also contains a timestamp (unix time) and a Readable TimeStamp (GMT+0).

* A device might contain multiple Scans.
* A Scan might contain multiple sets of process lists, but it’s usually one unique process list per Scan.
* A Scan is always limited to one device.
* A Scan could potentially only have one process entry.
* A Device might have multiple iOS Versions, when it has been updated during the time period.  
    
  The data set contains the following columns.



## **Task**

Your task is to analyze the provided data for anomalies and visualize the data set.  
  
Filename: pseudoProcess.csv  
Download URL:  
<https://drive.google.com/file/d/1D71ScZ9IQVDDB1ZGR9_fy6_UsUCi8qT2/view?usp=sharing>

Here is a couple of Questions that can help with the task:

* How many Devices are in the data set?
* How many Unique Scans are in the data?
* How many different iOS Versions are in the data set?
* How many Scans does a device have?
* Does iOS have a consistent naming scheme for processes?

Tools  
You are free to use whatever tools you like to work with to do the analysis. Please list all the tools that you used during the analysis. If you write code - please hand in the source code.  
 It’s not necessary to write code, but the job will contain quite a lot of prototyping and scripting. So showing your coding skills might be an advantage.  
  
 Final Remark  
This take home task is not about finding one exact process name that is malware. Its not even clear if malware is in the data set. The task is about showing your analysis skills and how you are approaching an unknown problem. A lot of our work is finding a needle in a haystack, not knowing if the needle is there or not.