ALGO TO FIND IOT RELATED IPs & QUERIES

Related ipynb file:
https://github.com/shahifaqeer/oit-dns/blob/master/feature_extraction/1hr-night-analysis.ipynb

How to find IPs that are IoTs in our dataset?

- Find IP, queries that are background data
- If srcip only producing background data during late night hours => classify as IoT
- If number of unique queries < THRESH and number of unique 1st level domain queries < THRESH2 => classify as IoT (?)

How to find (IP, query) that is background data?

- groupby query
- groupby srcip
- enumerate 'time' list
- diff 'time' list
- count, mean, std of diff time list vs srcip
- if for query Q, (count, mean, std) for most srcip are "similar" => Q, srcip are background data

Why use full query and not 1st/2nd level domains?

- summing stats for multiple queries => can't perform above time baded IoT/bg traffic filtering properly as counts, means etc will mess up due to a mix of unrelated query timings
- summing 1/2nd level domain only useful for absolute filter (eg: unique queries per IP to 1st level domain < THRESH etc.)

Summary

- Time series are too sparse to be interpreted meaningfuly by periodogram
- we get high point around freq ~ 0.1 when using sampling rate of 2 Hz
- but fitbit timeseries should have freq of 1/900 ~ 0.001
- huge time gaps and sparsity make frequency analysis cumbersome => can't separate bg data series
- any other way?