de_test

April 25, 2021

0.1 This File details the stages of data Extraction, Transformation, and Loading of a random 'csv file'

For more information about this file and related files, please see the Github Project Repository:

https://github.com/schwill2018/Data_Formatting

0.1.1 Import Necessary Packages & Load 'csv' file

[1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option("display.max_columns",100)

#Load File
test = pd.read_csv(r'C:/Users/schne/Desktop/projects/Tests/query-impala-111323.
 ↪csv')
```

0.1.2 Initial Data Observation

[2]:
```
# Preview file
test.head()
```

[2]:
```
           ts                                   uid  \
0  1609425630  63764947-3550-4c9c-a0ae-8f04cd5edfab
1  1609425634  95f1575f-06e7-47bf-aee5-ea36eb3bcba7
2  1609425640  5b925e8d-fb65-4504-87b1-f08bab9001fb
3  1609425640  a84e8319-ee1d-43cb-ad42-77d891bfcc0a
4  1609425641  18e0c945-24e0-462c-bcb8-5f030eaf2f45

                       acssessionid              cat_name  configversionid  \
0  phddcpsn2ise/377908621/219655882  CISE_Failed_Attempts            197.0
1  hfwdcpsn2ise/378174937/206479594  CISE_Failed_Attempts            187.0
2  phddcpsn2ise/377908621/219655883  CISE_Failed_Attempts            197.0
3                               NaN  CISE_Failed_Attempts              NaN
4  phddcpsn3ise/378178052/224827422  CISE_Failed_Attempts            171.0

  destinationipaddress  destinationipaddress_int  destinationport  \
```

```
0        10.198.252.93               180812893.0           1645.0
1        10.199.252.93               180878429.0           1645.0
2        10.198.252.93               180812893.0           1645.0
3                  NaN                       NaN              NaN
4        10.198.252.94               180812894.0           1645.0

  device_ip_address  device_ip_address_int  device_port  event_id  \
0     10.206.128.33            181305377.0      63722.0    5400.0
1     10.206.128.33            181305377.0      63722.0    5400.0
2       10.208.9.80            181406032.0      59665.0    5400.0
3               NaN                    NaN          NaN       NaN
4     10.206.128.33            181305377.0      63722.0    5400.0

                             event_text  \
0  Failed-Attempt: Authentication failed
1  Failed-Attempt: Authentication failed
2  Failed-Attempt: Authentication failed
3                                   NaN
4  Failed-Attempt: Authentication failed

                                      failurereason          host     msg_id  \
0  22056 Subject not found in the applicable iden…  phddcpsn2ise  174806944
1  22056 Subject not found in the applicable iden…  hfwdcpsn2ise  166833862
2  22056 Subject not found in the applicable iden…  phddcpsn2ise  174806945
3                                              NaN  phddcpsn2ise  174806945
4  22056 Subject not found in the applicable iden…  phddcpsn3ise  176078137

  nas_ip_address  nas_ip_address_int  \
0  10.206.128.33         181305377.0
1  10.206.128.33         181305377.0
2    10.208.9.80         181406032.0
3            NaN                 NaN
4  10.206.128.33         181305377.0

                        networkdevicegroups networkdevicename protocol  \
0                                       NaN       phdpb34b1cs   Radius
1                                       NaN       phdpb34b1cs   Radius
2                                       NaN       hebpb25a1cs   Radius
3  Device Type > All Device Types > Switches               NaN      NaN
4                                       NaN       phdpb34b1cs   Radius

                                              raw  requestlatency  seg_num  \
0  CISE_Failed_Attempts 0174806944 2 0 2020-12-31…             4.0        0
1  CISE_Failed_Attempts 0166833862 2 0 2020-12-31…             4.0        0
2  CISE_Failed_Attempts 0174806945 2 0 2020-12-31…             8.0        0
3  CISE_Failed_Attempts 0174806945 2 1  Step=2206…             NaN        1
4  CISE_Failed_Attempts 0176078137 2 0 2020-12-31…             6.0        0
```

```
  service_type sev_label  severity  total_seg   user_name    yr  mo  dy  \
0     16777216    NOTICE         5          2  s~f5health  2020  12  31
1     16777216    NOTICE         5          2  s~f5health  2020  12  31
2     16777216    NOTICE         5          2  s~f5health  2020  12  31
3          NaN       NaN         5          2         NaN  2020  12  31
4     16777216    NOTICE         5          2  s~f5health  2020  12  31

   archived
0      True
1      True
2      True
3      True
4      True
```

[3]:
```
test['raw'][3]
```

[3]:
```
'CISE_Failed_Attempts 0174806945 2 1  Step=22061, Step=11003,
SelectedAuthenticationIdentityStores=Internal Users,
NetworkDeviceGroups=Location#All Locations#HEB#PB2#2nd floor,
NetworkDeviceGroups=Device Type#All Device Types#Switches,
CPMSessionID=0ac6fc5d/k79oCTzhIk9nB_TXZa5bVOEEBjCf6ZU3BL08jNs84I,
ISEPolicySetName=Switch_Policy, AllowedProtocolMatchedRule=Default,
IdentitySelectionMatchedRule=Default, StepData=5= DEVICE.Device Type,
StepData=6= Radius.Service-Type, StepData=7= Normalised Radius.RadiusFlowType,
StepData=11=Internal Users, Network Device Profile=Cisco, Location=Location#All
Locations#HEB#PB2#2nd floor, Device Type=Device Type#All Device Types#Switches,
Response={RadiusPacketType=AccessReject; AuthenticationResult=UnknownUser; },'
```

0.1.3 Get Detailed Information about dataset via descriptive statistics and summary data

[4]:
```
print(test.describe())
print(test.nunique(axis = 0, dropna = True).sort_values(ascending=False))
print(test.isnull().sum(axis=0).sort_values(ascending=False))
```

```
                 ts  configversionid  destinationipaddress_int  \
count  1.000000e+03       758.000000              7.190000e+02
mean   1.609639e+09       189.432718              1.808511e+08
std    2.430840e+05        10.339678              3.233832e+04
min    1.609423e+09       171.000000              1.808129e+08
25%    1.609425e+09       187.000000              1.808129e+08
50%    1.609426e+09       197.000000              1.808784e+08
75%    1.609915e+09       197.000000              1.808784e+08
max    1.609916e+09       197.000000              1.808784e+08

       destinationport  device_ip_address_int   device_port      event_id  \
count       719.000000           7.190000e+02    705.000000    775.000000
```

```
mean       1645.002782           1.813562e+08  60917.546099   9203.974194
std           0.052705           5.036015e+04   7041.420663  14850.410699
min        1645.000000           1.813054e+08   1645.000000   5200.000000
25%        1645.000000           1.813054e+08  59665.000000   5400.000000
50%        1645.000000           1.814060e+08  59665.000000   5400.000000
75%        1645.000000           1.814060e+08  63722.000000   5400.000000
max        1646.000000           1.814060e+08  63722.000000  80002.000000

              msg_id  nas_ip_address_int  requestlatency      seg_num  \
count  1.000000e+03        7.190000e+02      718.000000  1000.000000
mean   1.720459e+08        1.813562e+08        9.179666     0.296000
std    3.318329e+06        5.036015e+04       18.176257     0.746285
min    1.668333e+08        1.813054e+08        4.000000     0.000000
25%    1.710520e+08        1.813054e+08        6.000000     0.000000
50%    1.711595e+08        1.814060e+08        6.000000     0.000000
75%    1.749140e+08        1.814060e+08        7.000000     0.000000
max    1.761930e+08        1.814060e+08      235.000000     7.000000

          severity    total_seg           yr           mo          dy
count  1000.00000  1000.000000  1000.000000  1000.000000  1000.00000
mean      5.02900     2.174000  2020.436000     7.204000    20.10000
std       0.17375     1.023118     0.496135     5.457487    12.40338
min       4.00000     1.000000  2020.000000     1.000000     6.00000
25%       5.00000     2.000000  2020.000000     1.000000     6.00000
50%       5.00000     2.000000  2020.000000    12.000000    31.00000
75%       5.00000     2.000000  2021.000000    12.000000    31.00000
max       6.00000     8.000000  2021.000000    12.000000    31.00000
raw                         1000
uid                         1000
msg_id                       853
ts                           782
acssessionid                 707
requestlatency                30
event_id                      11
event_text                    11
user_name                      9
seg_num                        8
cat_name                       6
total_seg                      6
destinationipaddress           5
destinationipaddress_int       5
host                           5
device_port                    4
configversionid                3
severity                       3
sev_label                      3
service_type                   3
dy                             2
```

```
yr                              2
mo                              2
nas_ip_address                  2
networkdevicename               2
nas_ip_address_int              2
device_ip_address_int           2
device_ip_address               2
destinationport                 2
protocol                        1
networkdevicegroups             1
failurereason                   1
archived                        1
dtype: int64
networkdevicegroups           805
failurereason                 297
device_port                   295
acssessionid                  293
service_type                  283
requestlatency                282
protocol                      282
networkdevicename             282
nas_ip_address_int            281
user_name                     281
nas_ip_address                281
device_ip_address_int         281
device_ip_address             281
destinationport               281
destinationipaddress_int      281
destinationipaddress          281
configversionid               242
event_id                      225
sev_label                     225
event_text                    225
mo                              0
yr                              0
total_seg                       0
dy                              0
severity                        0
ts                              0
seg_num                         0
raw                             0
uid                             0
msg_id                          0
host                            0
cat_name                        0
archived                        0
dtype: int64
```

[5]:
```
#Get shape of DF
test.shape
```

[5]:
```
(1000, 33)
```

[6]:
```
#Get shape of DF with NA's removed (by x-axis)
test.dropna(axis=0,how='any').shape
```

[6]:
```
(0, 33)
```

[7]:
```
#Get shape of DF with NA's removed (by y-axis)
print(test.dropna(axis=1,how= 'any').shape)
col_drop = test.dropna(axis=1,how= 'any')
```

```
(1000, 13)
```

Since dropping NA values by x-axis returns an empty dataframe, we can drop NA's by y-axis (or "columns") for comparison to original dataset.

We can also subset the data to drop NA's only where null values are found in specified fields.

Both of these null value transformations are detailed further below

0.1.4 Before we can address null values, we need to TRANSFORM the data into a workable database format

[8]:
```
#Format Date Data
test = test.rename(columns = {'yr':'year','mo':'month','dy':'day'})

dates=pd.to_datetime(test['ts'], unit = 's').dt.date
times=pd.to_datetime(test['ts'], unit = 's').dt.time

test['date'] = dates
test['time'] = times
test.tail()
```

[8]:

	ts	uid	\
995	1609913074	4675b6b9-8fa8-4e65-8e2c-d6884a6d86f5	
996	1609912812	ba2faeda-8bbd-4a9a-ad2e-992772cbf60c	
997	1609912813	370a9b8d-cf0f-48e8-a97e-44983eb44180	
998	1609912822	088a6c15-b0f0-4402-b26c-f72190e6a25d	
999	1609912847	3b7be59a-3644-43ef-b422-5ab845f585f0	

	acssessionid	cat_name	configversionid	\
995	phddcpsn2ise/377908621/219762316	CISE_Failed_Attempts	197.0	
996	NaN	CISE_Failed_Attempts	NaN	
997	hfwdcpsn2ise/378174937/206585954	CISE_Failed_Attempts	187.0	

```
998  phddcpsn2ise/377908621/219762259  CISE_Failed_Attempts            197.0
999  hfwdcpsn3ise/377906992/213902041  CISE_Failed_Attempts            197.0

    destinationipaddress  destinationipaddress_int  destinationport  \
995        10.198.252.93               180812893.0           1645.0
996                  NaN                       NaN              NaN
997        10.199.252.93               180878429.0           1645.0
998        10.198.252.93               180812893.0           1645.0
999        10.199.252.94               180878430.0           1645.0

    device_ip_address  device_ip_address_int  device_port  event_id  \
995       10.208.9.80            181406032.0      59665.0    5400.0
996               NaN                    NaN          NaN       NaN
997     10.206.128.33            181305377.0      63722.0    5400.0
998     10.206.128.33            181305377.0      63722.0    5400.0
999       10.208.9.80            181406032.0      59665.0    5400.0

                              event_text  \
995  Failed-Attempt: Authentication failed
996                                 NaN
997  Failed-Attempt: Authentication failed
998  Failed-Attempt: Authentication failed
999  Failed-Attempt: Authentication failed

                                        failurereason          host  \
995  22056 Subject not found in the applicable iden…  phddcpsn2ise
996                                               NaN  phddcpsn2ise
997  22056 Subject not found in the applicable iden…  hfwdcpsn2ise
998  22056 Subject not found in the applicable iden…  phddcpsn2ise
999  22056 Subject not found in the applicable iden…  hfwdcpsn3ise

       msg_id nas_ip_address  nas_ip_address_int  \
995  174913506    10.208.9.80         181406032.0
996  174913448            NaN                 NaN
997  166949326  10.206.128.33         181305377.0
998  174913450  10.206.128.33         181305377.0
999  171158814    10.208.9.80         181406032.0

                         networkdevicegroups networkdevicename protocol  \
995                                      NaN       hebpb25a1cs   Radius
996  Device Type > All Device Types > Switches               NaN      NaN
997                                      NaN       phdpb34b1cs   Radius
998                                      NaN       phdpb34b1cs   Radius
999                                      NaN       hebpb25a1cs   Radius

                                                raw  requestlatency  \
995  CISE_Failed_Attempts 0174913506 2 0 2021-01-06…             7.0
```

```
996  CISE_Failed_Attempts 0174913448 2 1  Step=2206…             NaN
997  CISE_Failed_Attempts 0166949326 2 0 2021-01-06…             6.0
998  CISE_Failed_Attempts 0174913450 2 0 2021-01-06…             7.0
999  CISE_Failed_Attempts 0171158814 2 0 2021-01-06…             6.0

     seg_num service_type sev_label  severity  total_seg   user_name  year  \
995        0     16777216    NOTICE         5          2  s~f5health  2021
996        1          NaN       NaN         5          2         NaN  2021
997        0     16777216    NOTICE         5          2  s~f5health  2021
998        0     16777216    NOTICE         5          2  s~f5health  2021
999        0     16777216    NOTICE         5          2  s~f5health  2021

     month  day  archived        date      time
995      1    6      True  2021-01-06  06:04:34
996      1    6      True  2021-01-06  06:00:12
997      1    6      True  2021-01-06  06:00:13
998      1    6      True  2021-01-06  06:00:22
999      1    6      True  2021-01-06  06:00:47
```

[9]:
```
test['networkdevicegroups'].value_counts(dropna=False)
```

[9]:
```
NaN                                          805
Device Type > All Device Types > Switches    195
Name: networkdevicegroups, dtype: int64
```

The column 'networkdevicegroups' contains two unique values, one of the values being null.

The other value appears to designate the device type as "Switches".

However, the value "Switches" is nested in device type subgroups.

0.1.5 The next line of code aims to solve for this format by applying Regex syntax in order to separate subgroups from "Device Type > All Device Types > Switches".

[10]:
```
#Split field 'netowrkdevicegroups' on the " > " condition, using Regex format
ndg_splt=test['networkdevicegroups'].str.split(r'.\>.',expand=True)
ndg_splt.head()
```

[10]:
```
              0                 1         2
0           NaN               NaN       NaN
1           NaN               NaN       NaN
2           NaN               NaN       NaN
3  Device Type  All Device Types  Switches
4           NaN               NaN       NaN
```

Currently, there is only one set of 'real' values in the 'networkdevicegroups' column.

However, if these groups were ever to vary with in the input of future data, it would be beneficial to view these subgroups within the frame of the whole dataset

The next line of code splits these "subgroups", and assigns each subgroup to a newly assigned column in the original dataset

[11]:
```
#Function to create  'c' new 'networkdevicegroup' subgroup fields.
for c in ndg_splt.columns:
    new_col='nwdev_subgrp{}'.format(c)
    test[str(new_col)] = ndg_splt[[int(c)]]

#Assign columns to list
cols = test.columns.tolist()

#Call as a series to determine by index-integer-location where to arrange fields
print(pd.Series(cols))

#Rearrage columns to place new 'nwdev_subgrp{}s' adjacent to origianl field
cols_nwd = cols[18:19] + cols[35:len(cols)]
nwdgrp=test[cols_nwd]
nwdgrp.head()
```

```
0                          ts
1                         uid
2                acssessionid
3                    cat_name
4              configversionid
5         destinationipaddress
6     destinationipaddress_int
7               destinationport
8            device_ip_address
9        device_ip_address_int
10                device_port
11                   event_id
12                 event_text
13              failurereason
14                       host
15                     msg_id
16             nas_ip_address
17         nas_ip_address_int
18        networkdevicegroups
19          networkdevicename
20                   protocol
21                        raw
22             requestlatency
23                    seg_num
24               service_type
```

```
25                   sev_label
26                    severity
27                   total_seg
28                   user_name
29                        year
30                       month
31                         day
32                    archived
33                        date
34                        time
35               nwdev_subgrp0
36               nwdev_subgrp1
37               nwdev_subgrp2
dtype: object
```

[11]:

	networkdevicegroups	nwdev_subgrp0	nwdev_subgrp1	nwdev_subgrp2
0	NaN	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN
3	Device Type > All Device Types > Switches	Device Type	All Device Types	Switches
4	NaN	NaN	NaN	NaN

0.1.6 The same transformation is applied to similar fields...

[12]:
```
# 'event_text' transformation
ev_txt=test['event_text'].str.split(r'\: ',expand=True)

#observe data before creating new field assignments
ev_txt.value_counts()
```

[12]:
```
0                     1
Failed-Attempt        Authentication failed
703
Passed-Authentication  Authentication succeeded
14
System-Stats          ISE Process Health
12
                      ISE Utilization
10
 Profiler             Profiler EndPoint profiling event occurred
```

```
9
System-Stats            ISE Counters
8
 AD-Connector           Trusted domain discovered
7
                        DC record cached
6
                        GC record cached
4
 RADIUS                 NAS sends RADIUS accounting update messages too
frequently              1
RADIUS                  NAS conducted several failed authentications of the same
scenario      1
dtype: int64
```

[13]:
```
#Create New fields '{}_status' & '{}_msg' for split fields from 'event_text'
new_col='evnt_status'
new_col1='evnt_msg'
test[str(new_col)] = ev_txt[[0]]
test[str(new_col1)] = ev_txt[[1]]

#Assign columns to list
cols = test.columns.tolist()

#Call as a series to determine by index-integer-location where to arrange fields
#print(pd.Series(cols))

#Rearrage columns to place new 'ev_txt{}s' adjacent to origianl field
cols_event = cols[11:13] + cols[38:len(cols)]
test[cols_event].head(3)
```

[13]:
```
   event_id                              event_text     evnt_status  \
0    5400.0  Failed-Attempt: Authentication failed  Failed-Attempt
1    5400.0  Failed-Attempt: Authentication failed  Failed-Attempt
2    5400.0  Failed-Attempt: Authentication failed  Failed-Attempt

                evnt_msg
0  Authentication failed
1  Authentication failed
2  Authentication failed
```

[14]:
```
# 'failurereason' transformation
print(test['failurereason'].value_counts())
fails=test['failurereason'].str.split(r'(?<=\d)\ ',expand=True)
test['failurenumb']=fails[[0]]
test['failurereason']= fails[[1]]
```

```
#Assign columns to list
cols = test.columns.tolist()

#Call as a series to determine by index-integer-location where to arrange fields
#print(pd.Series(cols))

#Rearrage columns to place new 'failurereason/numb{}s' adjacent to origianl␣
 ↪field
cols_fails = cols[13:14] + cols[40:len(cols)]
test[cols_fails].head(3)
```

```
22056 Subject not found in the applicable identity store(s)    703
Name: failurereason, dtype: int64
```

[14]:

	failurereason	failurenumb
0	Subject not found in the applicable identity s…	22056
1	Subject not found in the applicable identity s…	22056
2	Subject not found in the applicable identity s…	22056

[15]:
```
#'acssessionid' transformation
acsid=test['acssessionid'].str.split(r'\/',expand=True)

#Function to create  'c' new 'acssessionid' subgroup fields.
for c in acsid.columns:
    new_col='acssessionid_{}'.format(c)
    test[str(new_col)] = acsid[[int(c)]]
```

[16]:
```
#Assign columns to list
cols = test.columns.tolist()

#Call as a series to determine by index-integer-location where to arrange fields
#print(pd.Series(cols))

#Rearrage columns to place new 'acssessionid{}s' adjacent to origianl field
cols_acsid = cols[2:3] + cols[40:len(cols)]
test[cols_acsid].head(3)
```

[16]:

	acssessionid	failurenumb	acssessionid_0	acssessionid_1	acssessionid_2
0	phddcpsn2ise/377908621/219655882	22056	phddcpsn2ise	377908621	219655882
1	hfwdcpsn2ise/378174937/206479594	22056	hfwdcpsn2ise	378174937	206479594
2	phddcpsn2ise/377908621/219655883	22056	phddcpsn2ise	377908621	219655883

0.2 With current date formatting completed, let's take a look at the two dataframes to see if we can draw some insights about the data and it's null values

0.2.1 Regenerate preview of data

[17]:
```
test.head()
```

[17]:

	ts	uid	acssessionid	cat_name	configversionid	destinationipaddress	destinationipaddress_int	destinationport	device_ip_address	device_ip_address_int	device_port	event_id	event_text	failurereason	host	msg_id
0	1609425630	63764947-3550-4c9c-a0ae-8f04cd5edfab	phddcpsn2ise/377908621/219655882	CISE_Failed_Attempts	197.0	10.198.252.93	180812893.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn2ise	174806944
1	1609425634	95f1575f-06e7-47bf-aee5-ea36eb3bcba7	hfwdcpsn2ise/378174937/206479594	CISE_Failed_Attempts	187.0	10.199.252.93	180878429.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	hfwdcpsn2ise	166833862
2	1609425640	5b925e8d-fb65-4504-87b1-f08bab9001fb	phddcpsn2ise/377908621/219655883	CISE_Failed_Attempts	197.0	10.198.252.93	180812893.0	1645.0	10.208.9.80	181406032.0	59665.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn2ise	174806945
3	1609425640	a84e8319-ee1d-43cb-ad42-77d891bfcc0a	NaN	CISE_Failed_Attempts	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	phddcpsn2ise	174806945
4	1609425641	18e0c945-24e0-462c-bcb8-5f030eaf2f45	phddcpsn3ise/378178052/224827422	CISE_Failed_Attempts	171.0	10.198.252.94	180812894.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn2ise	174806946

```
4  Subject not found in the applicable identity s…  phddcpsn3ise  176078137

  nas_ip_address  nas_ip_address_int  \
0  10.206.128.33         181305377.0
1  10.206.128.33         181305377.0
2    10.208.9.80         181406032.0
3            NaN                 NaN
4  10.206.128.33         181305377.0

                        networkdevicegroups networkdevicename protocol  \
0                                       NaN       phdpb34b1cs   Radius
1                                       NaN       phdpb34b1cs   Radius
2                                       NaN       hebpb25a1cs   Radius
3  Device Type > All Device Types > Switches               NaN      NaN
4                                       NaN       phdpb34b1cs   Radius

                                              raw  requestlatency  seg_num  \
0  CISE_Failed_Attempts 0174806944 2 0 2020-12-31…             4.0        0
1  CISE_Failed_Attempts 0166833862 2 0 2020-12-31…             4.0        0
2  CISE_Failed_Attempts 0174806945 2 0 2020-12-31…             8.0        0
3  CISE_Failed_Attempts 0174806945 2 1  Step=2206…             NaN        1
4  CISE_Failed_Attempts 0176078137 2 0 2020-12-31…             6.0        0

  service_type sev_label  severity  total_seg   user_name  year  month  day  \
0     16777216    NOTICE         5          2  s~f5health  2020     12   31
1     16777216    NOTICE         5          2  s~f5health  2020     12   31
2     16777216    NOTICE         5          2  s~f5health  2020     12   31
3          NaN       NaN         5          2         NaN  2020     12   31
4     16777216    NOTICE         5          2  s~f5health  2020     12   31

   archived        date      time nwdev_subgrp0     nwdev_subgrp1  \
0      True  2020-12-31  14:40:30           NaN               NaN
1      True  2020-12-31  14:40:34           NaN               NaN
2      True  2020-12-31  14:40:40           NaN               NaN
3      True  2020-12-31  14:40:40   Device Type  All Device Types
4      True  2020-12-31  14:40:41           NaN               NaN

  nwdev_subgrp2     evnt_status               evnt_msg failurenumb  \
0           NaN  Failed-Attempt  Authentication failed       22056
1           NaN  Failed-Attempt  Authentication failed       22056
2           NaN  Failed-Attempt  Authentication failed       22056
3      Switches             NaN                   NaN         NaN
4           NaN  Failed-Attempt  Authentication failed       22056

  acssessionid_0 acssessionid_1 acssessionid_2
0   phddcpsn2ise      377908621      219655882
1   hfwdcpsn2ise      378174937      206479594
```

```
2   phddcpsn2ise      377908621      219655883
3            NaN            NaN            NaN
4   phddcpsn3ise      378178052      224827422
```

Unique Values Exlcuding Nulls

[18]:
```
test.nunique(dropna=False).sort_values()
```

[18]:
```
archived                      1
year                          2
nwdev_subgrp1                 2
day                           2
protocol                      2
nwdev_subgrp2                 2
networkdevicegroups           2
failurenumb                   2
nwdev_subgrp0                 2
date                          2
failurereason                 2
month                         2
severity                      3
nas_ip_address_int            3
nas_ip_address                3
networkdevicename             3
device_ip_address_int         3
device_ip_address             3
destinationport               3
configversionid               4
service_type                  4
sev_label                     4
device_port                   5
host                          5
destinationipaddress_int      6
destinationipaddress          6
acssessionid_1                6
cat_name                      6
acssessionid_0                6
total_seg                     6
evnt_status                   8
seg_num                       8
user_name                    10
event_text                   12
evnt_msg                     12
event_id                     12
requestlatency               31
acssessionid                708
acssessionid_2              708
ts                          782
```

```
time                          782
msg_id                        853
uid                          1000
raw                          1000
dtype: int64
```

There are 1000 rows of data. There is also 1000 unique instances of class 'uid'.

These id's appear to be data entry id's, because the only other field with as many unique values as index count is 'raw' ##### 'raw' is a field with each data entry compressed into one value

[19]:
```
print(test['acssessionid_0'].value_counts())
print(test['acssessionid_0'].value_counts())
test['acssessionid'].value_counts()
```

```
phddcpsn3ise    151
phddcpsn2ise    142
hfwdcpsn3ise    139
hfwdcpsn1ise    138
hfwdcpsn2ise    137
Name: acssessionid_0, dtype: int64
phddcpsn3ise    151
phddcpsn2ise    142
hfwdcpsn3ise    139
hfwdcpsn1ise    138
hfwdcpsn2ise    137
Name: acssessionid_0, dtype: int64
```

[19]:
```
hfwdcpsn3ise/377906992/213902792    1
hfwdcpsn1ise/377906300/213985952    1
hfwdcpsn1ise/377906300/213986313    1
hfwdcpsn3ise/377906992/213795520    1
hfwdcpsn1ise/377906300/214092567    1
                                   ..
phddcpsn3ise/378178052/224826973    1
hfwdcpsn1ise/377906300/213986055    1
phddcpsn2ise/377908621/219762797    1
hfwdcpsn3ise/377906992/213795348    1
phddcpsn3ise/378178052/224827682    1
Name: acssessionid, Length: 707, dtype: int64
```

Looking at the formatted dataset, along with the unique values (both stated above) the distribution of the values of 'accessionid' appears to confirm a condition. Since 'acssessionid_0' = 'host', and since the count of unique values 'acssessionid_1' = count 'acssessionid_0' we can infer that 'accessionid_1' is strongly tied to 'host'.

Also the # of unique values of'acssessionid' = # of unique values of 'acssesionid_2'. This suggest that that 'acssessionid_2' is the generator of unique values for the field 'acssessionid'

0.2.2 Observing Nulls Values further...

Total sum of null values per field/column (sorted descending)

[20]:
```
nulls_sum=pd.DataFrame(test.isnull().sum().sort_values(ascending=False).
↪reset_index())
nulls_sum.head()
```

[20]:

	index	0
0	networkdevicegroups	805
1	nwdev_subgrp2	805
2	nwdev_subgrp1	805
3	nwdev_subgrp0	805
4	failurenumb	297

At first glance, we can see the column 'networkdevicegroups' has 800+ NA values.

Another observation: other fields containing NA values have similar sums of null values. This might be indicative that these fields' null results are related

0.2.3 Now comes the time to address the NA's. What do we do with them?

0.2.4 Some entries might still contain critical data even if the value is missing. But we also do not want to pass NaN values to the final file, otherwise that could cause issues for anyone else downstream accessing this data

First, we'll trying dropping NA/null values by column

[22]:
```
data_col_drop=test.dropna(axis=1,how='any')
data_col_drop.head()
```

[22]:

	ts	uid	cat_name	\
0	1609425630	63764947-3550-4c9c-a0ae-8f04cd5edfab	CISE_Failed_Attempts	
1	1609425634	95f1575f-06e7-47bf-aee5-ea36eb3bcba7	CISE_Failed_Attempts	
2	1609425640	5b925e8d-fb65-4504-87b1-f08bab9001fb	CISE_Failed_Attempts	
3	1609425640	a84e8319-ee1d-43cb-ad42-77d891bfcc0a	CISE_Failed_Attempts	
4	1609425641	18e0c945-24e0-462c-bcb8-5f030eaf2f45	CISE_Failed_Attempts	

	host	msg_id		raw	\
0	phddcpsn2ise	174806944	CISE_Failed_Attempts 0174806944 2 0 2020-12-31…		
1	hfwdcpsn2ise	166833862	CISE_Failed_Attempts 0166833862 2 0 2020-12-31…		
2	phddcpsn2ise	174806945	CISE_Failed_Attempts 0174806945 2 0 2020-12-31…		
3	phddcpsn2ise	174806945	CISE_Failed_Attempts 0174806945 2 1 Step=2206…		
4	phddcpsn3ise	176078137	CISE_Failed_Attempts 0176078137 2 0 2020-12-31…		

	seg_num	severity	total_seg	year	month	day	archived	date	\

```
0        0         5          2  2020     12   31      True  2020-12-31
1        0         5          2  2020     12   31      True  2020-12-31
2        0         5          2  2020     12   31      True  2020-12-31
3        1         5          2  2020     12   31      True  2020-12-31
4        0         5          2  2020     12   31      True  2020-12-31

       time
0  14:40:30
1  14:40:34
2  14:40:40
3  14:40:40
4  14:40:41
```

This dataset turned out nice. Null values are non-existent

However...

We lost more than half of the columns in the original dataset (before any data transformations were even applied!!)

0.2.5 Another option to address null values:

Remove null values by the x-axis, and since each row of data has at least 1 null value, we can subset the columns of nulls as a condition on a null drop of the data rows

[23]:
```
# Create a dataframe/array of all values that HAVE NA
for i in range(len(nulls_sum.index)):
    if int(nulls_sum.loc[[i],[0]].values) > 0 :
        pass
    else:
        nulls_sum = nulls_sum.drop(i,axis=0)


print(nulls_sum.head())
```

```
                 index    0
0  networkdevicegroups  805
1        nwdev_subgrp2  805
2        nwdev_subgrp1  805
3        nwdev_subgrp0  805
4          failurenumb  297
```

[24]:
```
#Drop rows where ALL null subsets are found
drop_lst=list(nulls_sum['index'])
na_drop=test.dropna(subset=drop_lst,how='all')
print(range(len(na_drop.index)))
print(range(len(na_drop.columns)))
#Fill missing values with "No Data"
data_row_drop=test.fillna('No Data')
```

```
data_row_drop.head()
```

```
range(0, 974)
range(0, 44)
```

[24]:

	ts	uid	acssessionid	cat_name	configversionid	destinationipaddress	destinationipaddress_int	destinationport	device_ip_address	device_ip_address_int	device_port	event_id	event_text	failurereason	host	msg_id
0	1609425630	63764947-3550-4c9c-a0ae-8f04cd5edfab	phddcpsn2ise/377908621/219655882	CISE_Failed_Attempts	197.0	10.198.252.93	180812893.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn2ise	174806944
1	1609425634	95f1575f-06e7-47bf-aee5-ea36eb3bcba7	hfwdcpsn2ise/378174937/206479594	CISE_Failed_Attempts	187.0	10.199.252.93	180878429.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	hfwdcpsn2ise	166833862
2	1609425640	5b925e8d-fb65-4504-87b1-f08bab9001fb	phddcpsn2ise/377908621/219655883	CISE_Failed_Attempts	197.0	10.198.252.93	180812893.0	1645.0	10.208.9.80	181406032.0	59665.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn2ise	174806945
3	1609425640	a84e8319-ee1d-43cb-ad42-77d891bfcc0a	No Data	CISE_Failed_Attempts	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	phddcpsn2ise	174806945
4	1609425641	18e0c945-24e0-462c-bcb8-5f030eaf2f45	phddcpsn3ise/378178052/224827422	CISE_Failed_Attempts	171.0	10.198.252.94	180812894.0	1645.0	10.206.128.33	181305377.0	63722.0	5400.0	Failed-Attempt: Authentication failed	Subject not found in the applicable identity s…	phddcpsn3ise	176078137

```
  nas_ip_address nas_ip_address_int  \
0  10.206.128.33        181305377.0
1  10.206.128.33        181305377.0
2    10.208.9.80        181406032.0
3        No Data            No Data
4  10.206.128.33        181305377.0

                        networkdevicegroups networkdevicename protocol  \
0                                  No Data       phdpb34b1cs   Radius
1                                  No Data       phdpb34b1cs   Radius
2                                  No Data       hebpb25a1cs   Radius
3  Device Type > All Device Types > Switches           No Data  No Data
4                                  No Data       phdpb34b1cs   Radius

                                                raw requestlatency  seg_num  \
0  CISE_Failed_Attempts 0174806944 2 0 2020-12-31…            4.0        0
1  CISE_Failed_Attempts 0166833862 2 0 2020-12-31…            4.0        0
2  CISE_Failed_Attempts 0174806945 2 0 2020-12-31…            8.0        0
3  CISE_Failed_Attempts 0174806945 2 1  Step=2206…        No Data        1
4  CISE_Failed_Attempts 0176078137 2 0 2020-12-31…            6.0        0

  service_type sev_label  severity  total_seg   user_name  year  month  day  \
0     16777216    NOTICE         5          2  s~f5health  2020     12   31
1     16777216    NOTICE         5          2  s~f5health  2020     12   31
2     16777216    NOTICE         5          2  s~f5health  2020     12   31
3      No Data   No Data         5          2     No Data  2020     12   31
4     16777216    NOTICE         5          2  s~f5health  2020     12   31

   archived        date      time nwdev_subgrp0     nwdev_subgrp1  \
0      True  2020-12-31  14:40:30       No Data           No Data
1      True  2020-12-31  14:40:34       No Data           No Data
2      True  2020-12-31  14:40:40       No Data           No Data
3      True  2020-12-31  14:40:40   Device Type  All Device Types
4      True  2020-12-31  14:40:41       No Data           No Data

  nwdev_subgrp2     evnt_status               evnt_msg failurenumb  \
0       No Data  Failed-Attempt  Authentication failed       22056
1       No Data  Failed-Attempt  Authentication failed       22056
2       No Data  Failed-Attempt  Authentication failed       22056
3      Switches         No Data                No Data     No Data
4       No Data  Failed-Attempt  Authentication failed       22056

  acssessionid_0 acssessionid_1 acssessionid_2
0   phddcpsn2ise      377908621      219655882
1   hfwdcpsn2ise      378174937      206479594
2   phddcpsn2ise      377908621      219655883
3        No Data        No Data        No Data
```

```
4    phddcpsn3ise       378178052       224827422
```

0.2.6 Now that the data has been extracted and transformed, we will load this data back into the folder of the original data file.

Remove Excess Columns from Data (with nulls dropped on the x-axis)

[25]:
```
# Drop orignalcolumns after split from orignal formatting
data_row_drop=data_row_drop.
 ↪drop(columns=['acssessionid','event_text','year','month','date'])
```

0.3 Load Data

0.3.1 The data was then loaded back into the respective folder of the original extraction location, as type 'csv' file.

0.3.2 The data can be formatted to upload in many formats, including but not limited to:

>'xlsx'

>'sql'

>'html'

>'json

[26]:
```
data_row_drop.to_csv(r'C:/Users/schne/Desktop/projects/Tests/data_row_drop.csv')
```

[27]:
```
data_col_drop.to_csv(r'C:/Users/schne/Desktop/projects/Tests/data_col_drop.csv')
```

[ ]:

[ ]: