

```
[22]: import pandas as pd
import numpy as np
pd.set_option('display.max_rows', 500)

#import or 'read' the csv file of the dataset
df_dpibdata = pd.read_csv('device_performance_index.csv')
df_dpibdata
```

Out [22]:

	Client_ID	Office_ID	BSOD_count	Hard_reset_count	Boot_Speed	Logon_Duration	CPU_Usage	Memory_Usage	System_Free_Space	
	0	1	88	0	0	9824	8520	0.005	0.000	180269825344
	1	1	124	0	11	204075	34036	0.001	0.000	190256279562
	2	1	172	1	21	51930	15774	0.000	0.000	198199214080
	3	1	732	0	0	167701	3633	0.000	0.000	198328692736
	4	1	172	0	1	235541	19340	0.009	0.000	177966186496

	65164	9	283	0	0	8136	18113	0.002	0.114	21484277760
	65165	9	636	0	0	30709	70357	0.000	0.000	377559977984
	65166	9	626	0	0	10907	40337	0.020	0.000	156750188544
	65167	9	124	0	0	12038	23904	0.004	0.000	145551847424
	65168	9	183	0	0	32162	40562	0.004	0.000	383183306752

65169 rows x 9 columns

In [23]:

```
#function to calculate the Xmin and Xmax takning into account the 2 and 98 percentiles
def xMin_xMax_calculator (dataFrame):

    xMinxMax = dataFrame.loc[:, ~dataFrame.columns.isin(['Client_ID', 'Office_ID'])].quantile([0.02, 0.98])

    return xMinxMax

resultado => xMin_xMax_calculator(df_dp1data)
resultado
```

Out [23]:

	BSOD_count	Hard_reset_count	Boot_Speed	Logon_Duration	CPU_Usage	Memory_Usage	System_Free_Space
0.02	0.0	0.0	4624.00	3708.36	0.000	0.00000	1.714007e+09

65169 rows × 9 columns

In [23]:

```
#function to calculate the xmin and xmax taking into account the 2 and 98 percentiles
def xMin_xMax_calculator (dataFrame):

    #iterate through the columns and normalize each value according to the given formula
    for column in dfCopy.columns[2:11]:
        x = ((dfCopy[column]-resultado[column])[0.02])/((resultado[column])[0.98]-resultado[column])[0.02]))
        #because both xmin and xmax (2 and 98 percentiles) for BSOD_count are 0, meaning that values greater than zero
        #are actually considered outliers, I replaced inf and nan values from Xnorm calculation.
        with pd.option_context('mode.use_inf_as_na', True):
            x.fillna(0, inplace=True)
        x = x.apply(lambda v: min(1, max(0, v)))
    dfCopy[column] = x

resultado = xMin_xMax_calculator(df_dpibdata)
resultado
```

Out [23]:

	BSOD_count	Hard_reset_count	Boot_Speed	Logon_Duration	CPU_Usage	Memory_Usage	System_Free_Space
0.02	0.0	0.0	4624.00	3708.36	0.000	0.00000	1.714007e+09
0.98	0.0	2.0	109884.68	96486.88	0.046	0.89164	4.471692e+11

In [24]:

```
#make a copy of the original data set, to perform operations with the data

dfCopy = df_dpibdata.copy()
#iterate through the columns and normalize each value according to the given formula
for column in dfCopy.columns[2:11]:
    x = ((dfCopy[column]-resultado[column])[0.02])/((resultado[column])[0.98]-resultado[column])[0.02]))
    #because both xmin and xmax (2 and 98 percentiles) for BSOD_count are 0, meaning that values greater than zero
    #are actually considered outliers, I replaced inf and nan values from Xnorm calculation.
    with pd.option_context('mode.use_inf_as_na', True):
        x.fillna(0, inplace=True)
    x = x.apply(lambda v: min(1, max(0, v)))
    dfCopy[column] = x

dfCopy
```

3	1	732	1	1.0	0.000000	1.000000	1.000000	1.000000	0.441379
4	1	172	1	1.0	0.000000	0.831535	0.804348	1.000000	0.395668
...
65164	9	283	1	1.0	0.966635	0.844758	0.956522	0.872146	0.044382
65165	9	636	1	1.0	0.782187	0.281715	1.000000	1.000000	0.843735
65166	9	626	1	1.0	0.840310	0.605246	0.565217	1.000000	0.348040
65167	9	124	1	1.0	0.929565	0.782348	0.913043	1.000000	0.322901
65168	9	183	1	1.0	0.738383	0.602921	0.913043	1.000000	0.856358

65169 rows x 9 columns

```
In [26]: #created a list with the names of the columns just to then, sum the values for each row, excluding the Client_Id and Office_Id
col_list = list(dfCopy)

del col_list[0:2]

dfCopy['DP1'] = (10/7) * dfCopy[col_list].sum(axis=1)

dfCopy

Client_ID  Office_ID  BSOD_count  Hard_reset_count  Boot_Speed  Logon_Duration  CPU_Usage  Memory_Usage  System_Free_Space  DPI
0          1          88           1              1.0      0.950599      0.948144      0.891304      1.000000      0.400839  0.844123
```

65169 rows × 9 columns

In [25]:

```
#iterate through the columns and flip the data over (1-v), except for the column System_Free_Space
for column in dfCopy.columns[2:8]:
    dfCopy[column] = dfCopy[column].apply(lambda v: 1-v)
dfCopy
```

65169 rows × 10 columns

In [27]:

I add the calculated DPI to the original dataframe just to show the raw data and the calculated DPI
df_dpiData['DPI'] = dfCopy['DPI']
df_dpiData

Out[27]:

	Client_ID	Office_ID	BSOD_count	Hard_reset_count	Boot_Speed	Logon_Duration	CPU_Usage	Memory_Usage	System_Free_Space	DPI
0	1	88	0	0	9824	8520	0.005	0.000	180269625344	8.844123
1	1	124	0	11	204075	34036	0.001	0.000	190256279552	5.820959
2	1	172	1	21	51930	15774	0.000	0.000	1981199214080	6.945196
3	1	732	0	0	167701	3633	0.000	0.000	198328692736	7.773399
4	1	172	0	1	235541	19340	0.009	0.000	177966186496	6.473643
...
65164	9	283	0	0	8136	18113	0.002	0.114	21484277760	8.120633
65165	9	636	0	0	30709	70357	0.000	0.000	377559977984	6.396623
65166	9	626	0	0	10907	40337	0.020	0.000	156750188544	7.798305
65167	9	124	0	0	12038	23904	0.004	0.000	145551847424	8.496939
65168	9	183	0	0	32162	40562	0.004	0.000	383183306752	8.729437

65169 rows × 10 columns

65169 rows × 9 columns

In [26]:

```
#I created a list with the names of the columns just to then, sum the values for each row, excluding the Client_ID and Office_ID
col_list = list(dfCopy)

del col_list[0:2]

dfCopy['DPI'] = (18/7) * dfCopy[col_list].sum(axis=1)

dfCopy
```

5	1431	8.542190	0.684110	8.735744	4.703036	9.962908	8.207064	9.002747	
6	1360	0.080344	0.877823	8.316343	4.370388	9.761934	7.471366	8.895051	
7	11870	7.916748	0.784605	8.013568	3.393475	9.940652	7.419211	8.455255	
8	9965	8.756074	0.784118	8.851732	4.547216	10.000000	8.373187	9.369694	
9	1809	7.909599	0.830721	8.170680	4.120610	9.635803	7.406426	8.429599	

Info between customers

- The client with the best DPI is number 8 with mean value of 8.76
- The client with the worse DPI is number 2 with a mean value of 7.58. This has the greatest std and the lowest min value detected. It has the greatest number of devices tested
- Client 3 has the lowest number of devices tested

In [29]:

```
#Data was grouped by Client Id and by Office Id to evaluate differences between offices for each client.
agglist = ['count', 'mean', 'std', 'median', 'min', 'max', q25, q75]
clientAndOffice_group = df_dpiData.groupby(['Client_ID', 'Office_ID'])['DPI'].agg(agglist)
clientAndOffice_group.head(100)

#Mean values for std exist because there are some offices that have only one device producing data, hence the std
```

65169 rows × 10 columns

In [27]:

```
# I add the calculated DPI to the original dataFrame just to show the raw data and the calculated DPI
df_dpibdata['DPI'] = dfCopy['DPI']
df_dpibdata
```

Out [27]:

	Client_ID	Office_ID	BSOD_count	Hard_reset_count	Boot_Speed	Logon_Duration	CPU_Usage	Memory_Usage	System_Free_Space	DPI	
	0	1	88	0	0	9824	8520	0.005	0.000	180269825344	8.844123
	1	1	124	0	11	204075	34036	0.001	0.000	190256279562	5.820959
	2	1	172	1	21	51930	15774	0.000	0.000	198199214080	6.945196
	3	1	732	0	0	167701	3633	0.000	0.000	198328692736	7.773399
	4	1	172	0	1	235541	19340	0.009	0.000	177966186496	6.473643

	65164	9	283	0	0	8136	18113	0.002	0.114	21484277760	8.120633
	65165	9	636	0	0	30709	70357	0.000	0.000	377559977984	8.396623
	65166	9	626	0	0	10907	40337	0.020	0.000	156750188544	7.798305
	65167	9	124	0	0	12038	23904	0.004	0.000	145551847424	8.496939
	65168	9	183	0	0	32162	40562	0.004	0.000	383183306752	8.729437

65169 rows × 10 columns

In [28]:

```
#I group the data by Client_ID and generate some statistics based on the calculated DPI

#first I create quantile functions to aggregate later
def q25(x):
    return x.quantile(0.25)
def q75(x):
    return x.quantile(0.75)
client_group = df_dpibdata.groupby(['Client_ID']).DPI.agg(['count', 'mean', 'std', 'median', 'min', 'max', 'q25, q75])
client_group
```

Out [28]:

	count	mean	std	median	min	max	q25	q75
Client_ID								
1	747	8.171885	0.749689	8.383374	5.258783	9.869012	7.715700	8.705456
2	27061	7.579961	0.958438	7.754950	2.381598	9.736532	7.013942	8.343630
3	77	8.616359	0.908401	8.514863	6.312591	9.956028	8.007548	9.384459
4	10849	8.414967	0.603874	8.541943	8.811278	9.985133	8.145786	8.812709
5	1431	8.542190	0.684110	8.735744	4.700336	9.962098	8.207064	9.002747
6	1360	8.080344	0.877923	8.316343	4.370388	9.761934	7.471366	8.859051
7	11870	7.916748	0.784605	8.013568	3.393475	9.940652	7.419211	8.455225
8	9965	8.756074	0.784118	8.851732	4.547216	10.000000	8.373187	9.369694
9	11809	7.908599	0.830721	8.170680	4.120810	9.635903	7.408426	8.425999

Info between customers

- The client with the best DPI is number 8 with mean value of 8.76
- The client with the worse DPI is number 2 with a mean value of 7.58. This has the greatest std and the lowest min value detected. It has the greatest number of devices tested
- Client 3 has the lowest number of devices tested

In [29]:

```
#data was grouped by Client_ID and by Office_ID to evaluate differences between offices for each client.
agglist = ['count', 'mean', 'std', 'median', 'min', 'max', 'q25, q75]
clientAndOffice_group = df_dpibdata.groupby(['Client_ID', 'Office_ID'])['DPI'].agg(agglist)
clientAndOffice_group.head(180)
```

NaN values for std exist because there are some offices that have only one device producing data, hence the std cannot be calculated.