In [1]:
```
import pandas as pd
```

In [2]:
```
data = pd.read_csv('C:\\Users\\91996\\Downloads\\01.Data Cleaning and Preprocessing.csv')
```

In [3]:
```
type(data)
```

Out[3]:
```
pandas.core.frame.DataFrame
```

In [4]:
```
data.info
```

Out[4]:
```
<bound method DataFrame.info of     Observation  Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  \
0      31-00:00    23.10    16.520     121.717  1177.607     169.805
1      31-01:00    27.60    16.810      79.022  1328.360     341.327
2      31-02:00    23.19    16.709      79.562  1329.407     239.161
3      31-03:00    23.60    16.478      81.011  1334.877     213.527
4      31-04:00    22.90    15.618      93.244  1334.168     243.131
..          ...      ...       ...         ...       ...         ...
319    10-16:00    23.75    12.667      93.450  1178.252     276.955
320     9-19:00    19.80    12.558      94.352  1184.119     297.071
321     9-20:00    23.01    12.550      90.842  1188.517     289.826
322     9-21:00    24.32    13.083      88.910  1192.879     318.006
323     9-22:00    25.75    13.417      85.451  1186.342     248.312

     T-upperExt-2   T-lowerExt-2    UCZAA  WhiteFlow-4   ...  SteamFlow-4
0          358.282         329.545  1.443      599.253   ...       67.122  \
1          351.050         329.067  1.549      537.201   ...       60.012
2          350.022         329.260  1.600      549.611   ...       61.304
3          350.938         331.142  1.604      623.362   ...       68.496
4          351.640         332.709    NaN      638.672   ...       70.022
..             ...             ...    ...          ...   ...          ...
319        347.286         310.970  1.523      513.956   ...       61.141
320        399.135         319.576  1.451      570.058   ...       67.667
321        373.633         314.591  1.457      549.306   ...       66.446
322        364.081         308.559  1.523      504.852   ...       61.054
323        356.289         310.482  1.474      497.375   ...       58.247

     Lower-HeatT-3  Upper-HeatT-3  ChipMass-4   WeakLiquorF   BlackFlow-2  \
0          329.432        303.099      175.964      1127.197      1319.039  \
1          330.823        304.879      163.202       665.975      1297.317
2          329.140        303.383      164.013       677.534      1327.072
3          328.875        302.254      181.487       767.853      1324.461
4          328.352        300.954      183.929       888.448      1343.424
..             ...            ...          ...           ...           ...
319        330.117        304.006      148.174      1027.201      1357.271
320        330.848        304.616      165.178       906.962      1311.177
321        330.226        304.686      160.841       887.125      1319.226
322        327.346        304.363      147.589       804.423      1320.225
323        328.092        304.093      144.218       828.328      1320.848

     WeakWashF   SteamHeatF-3   T-Top-Chips-4   SulphidityL-4
0      257.325         54.612          252.077             NaN
1      241.182         46.603          251.406           29.11
2      237.272         51.795          251.335             NaN
3      239.478         54.846          250.312           29.02
4      215.372         54.186          249.916           29.01
..         ...            ...              ...             ...
319    381.643         45.264          252.947           30.86
320     25.494         50.528          252.092           30.70
321      0.638         45.549          252.438             NaN
322      0.000         43.725          253.176           31.13
323      1.276         43.840          253.216             NaN

[324 rows x 23 columns]>
```

In [5]:
```
data.shape
```

Out[5]:
```
(324, 23)
```

In [6]:
```
data.describe()
```

Out[6]:

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	A
count	324.000000	319.000000	307.000000	308.000000	323.000000	322.000000	322.000000	299.000000	323.000000	1
mean	20.635370	14.347937	87.464456	1237.837614	258.164483	356.904295	324.020180	1.492010	591.732260	
std	3.070036	1.499095	7.995012	100.593735	87.987452	9.209290	7.621402	0.105923	67.016351	
min	12.170000	9.983000	68.645000	0.000000	0.000000	339.168000	284.633000	1.182000	405.111000	
25%	18.382500	13.358000	81.823000	1193.215250	213.527000	350.241250	321.420000	1.431500	540.989500	
50%	20.845000	14.308000	86.739000	1273.138500	271.792000	356.843000	325.669000	1.498000	592.895000	
75%	23.032500	15.517000	92.372000	1289.196000	321.680000	362.242250	329.175000	1.560500	639.480500	
max	27.600000	16.958000	121.717000	1351.240000	419.014000	399.135000	337.012000	1.747000	731.394000	

8 rows × 22 columns

In [8]:
```
data = data.drop_duplicates()
```

In [9]:
```
data
```

Out[9]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	NaN	638.672	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	
299	12-10:00	24.98	NaN	85.034	1278.345	368.564	357.723	321.387	NaN	520.365	...	
300	12-11:00	21.00	NaN	88.013	1307.722	278.842	357.438	323.757	NaN	553.070	...	
301	12-12:00	21.40	NaN	85.490	1255.986	273.484	361.365	322.689	NaN	590.199	...	
307	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	...	

301 rows × 23 columns

In [10]:
```
data.isnull()
```

Out[10]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	False	False	False	False	False	False	False	False	False	False	...	
1	False	False	False	False	False	False	False	False	False	False	...	
2	False	False	False	False	False	False	False	False	False	False	...	
3	False	False	False	False	False	False	False	False	False	False	...	
4	False	False	False	False	False	False	False	False	True	False	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	False	False	False	False	False	False	False	False	False	False	...	
299	False	False	True	False	False	False	False	False	True	False	...	
300	False	False	True	False	False	False	False	False	True	False	...	
301	False	False	True	False	False	False	False	False	True	False	...	
307	False	False	False	False	False	False	False	False	False	False	...	

301 rows × 23 columns

In [11]:
```
data.isnull().sum()
```

Out[11]:
```
Observation           0
Y-Kappa               0
ChipRate              4
BF-CMratio           14
BlowFlow             13
ChipLevel4            1
T-upperExt-2          1
T-lowerExt-2          1
UCZAA                24
WhiteFlow-4           1
AAWhiteSt-4         141
AA-Wood-4             1
ChipMoisture-4        1
SteamFlow-4           1
Lower-HeatT-3         1
Upper-HeatT-3         1
ChipMass-4            1
WeakLiquorF           1
BlackFlow-2           1
WeakWashF             1
SteamHeatF-3          1
T-Top-Chips-4         1
SulphidityL-4       141
dtype: int64
```

In [12]:
```
data.notnull()
```

Out[12]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	True	True	True	True	True	True	True	True	True	True	...	
1	True	True	True	True	True	True	True	True	True	True	...	
2	True	True	True	True	True	True	True	True	True	True	...	
3	True	True	True	True	True	True	True	True	True	True	...	
4	True	True	True	True	True	True	True	True	False	True	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	True	True	True	True	True	True	True	True	True	True	...	
299	True	True	False	True	True	True	True	True	False	True	...	
300	True	True	False	True	True	True	True	True	False	True	...	
301	True	True	False	True	True	True	True	True	False	True	...	
307	True	True	True	True	True	True	True	True	True	True	...	

301 rows × 23 columns

In [14]:
```
data.isnull().sum().sum()
```

Out[14]:
```
352
```

In [31]:
```
data2 = data.fillna(value=0)
data2
```

Out[31]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	0.000	638.672	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	
299	12-10:00	24.98	0.000	85.034	1278.345	368.564	357.723	321.387	0.000	520.365	...	
300	12-11:00	21.00	0.000	88.013	1307.722	278.842	357.438	323.757	0.000	553.070	...	
301	12-12:00	21.40	0.000	85.490	1255.986	273.484	361.365	322.689	0.000	590.199	...	
307	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	...	

301 rows × 23 columns

In [33]:
```
data2.isnull().sum().sum()
```

Out[33]:
```
0
```

In [34]:
```
data2
```

Out[34]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	0.000	638.672	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	
299	12-10:00	24.98	0.000	85.034	1278.345	368.564	357.723	321.387	0.000	520.365	...	
300	12-11:00	21.00	0.000	88.013	1307.722	278.842	357.438	323.757	0.000	553.070	...	
301	12-12:00	21.40	0.000	85.490	1255.986	273.484	361.365	322.689	0.000	590.199	...	
307	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	...	

301 rows × 23 columns

In [35]:
```
#filling null values with the next value
data3 = data.fillna(method = 'bfill')
data3
```

Out[35]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	1.436	638.672	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	
299	12-10:00	24.98	14.308	85.034	1278.345	368.564	357.723	321.387	1.522	520.365	...	
300	12-11:00	21.00	14.308	88.013	1307.722	278.842	357.438	323.757	1.522	553.070	...	
301	12-12:00	21.40	14.308	85.490	1255.986	273.484	361.365	322.689	1.522	590.199	...	
307	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	...	

301 rows × 23 columns

In [36]:
```
data4 = data.fillna(method = 'pad')
data4
```

Out[36]:

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	...	Stea
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	599.253	...	
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	...	
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	...	
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	...	
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	1.604	638.672	...	
...	...	...	...	...	...	...	...	...	...	...	...	
298	12-09:00	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	...	
299	12-10:00	24.98	15.167	85.034	1278.345	368.564	357.723	321.387	1.635	520.365	...	
300	12-11:00	21.00	15.167	88.013	1307.722	278.842	357.438	323.757	1.635	553.070	...	
301	12-12:00	21.40	15.167	85.490	1255.986	273.484	361.365	322.689	1.635	590.199	...	
307	31-05:00	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	...	

301 rows × 23 columns

In [37]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

In [38]:
```
#detect the outliners using IQR
data2.columns
```

Out[38]:
```
Index(['Observation', 'Y-Kappa', 'ChipRate', 'BF-CMratio', 'BlowFlow',
       'ChipLevel4 ', 'T-upperExt-2 ', 'T-lowerExt-2  ', 'UCZAA',
       'WhiteFlow-4 ', 'AAWhiteSt-4 ', 'AA-Wood-4  ', 'ChipMoisture-4 ',
       'SteamFlow-4 ', 'Lower-HeatT-3', 'Upper-HeatT-3 ', 'ChipMass-4 ',
       'WeakLiquorF ', 'BlackFlow-2 ', 'WeakWashF ', 'SteamHeatF-3 ',
       'T-Top-Chips-4 ', 'SulphidityL-4 '],
      dtype='object')
```

In [39]:
```
data2.drop(['Observation'], axis=1, inplace=True)
data2.columns
```

Out[39]:
```
Index(['Y-Kappa', 'ChipRate', 'BF-CMratio', 'BlowFlow', 'ChipLevel4 ',
       'T-upperExt-2 ', 'T-lowerExt-2  ', 'UCZAA', 'WhiteFlow-4 ',
       'AAWhiteSt-4 ', 'AA-Wood-4  ', 'ChipMoisture-4 ', 'SteamFlow-4 ',
       'Lower-HeatT-3', 'Upper-HeatT-3 ', 'ChipMass-4 ', 'WeakLiquorF ',
       'BlackFlow-2 ', 'WeakWashF ', 'SteamHeatF-3 ', 'T-Top-Chips-4 ',
       'SulphidityL-4 '],
      dtype='object')
```

In [40]:
```
Q1= data2.quantile(0.25)
Q3= data2.quantile(0.75)
IQR=Q3-Q1
print(IQR)
```

```
Y-Kappa               4.550
ChipRate              2.233
BF-CMratio           10.912
BlowFlow             96.766
ChipLevel4          105.868
T-upperExt-2         11.994
T-lowerExt-2          7.609
UCZAA                 0.152
WhiteFlow-4         100.098
AAWhiteSt-4           6.143
AA-Wood-4             1.486
ChipMoisture-4        2.186
SteamFlow-4           8.840
Lower-HeatT-3         8.585
Upper-HeatT-3         7.852
ChipMass-4           19.347
WeakLiquorF         180.613
BlackFlow-2         280.829
WeakWashF           267.219
SteamHeatF-3          6.903
T-Top-Chips-4         2.044
SulphidityL-4        30.420
dtype: float64
```

In [41]:
```
data2= data2[~((data2<(Q1-1.5*IQR))|(data2>(Q3+1.5*IQR))).any(axis=1)]
data2
```

Out[41]:

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	AAWhiteSt-4	...	Stea
1	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	537.201	6.076	...	
2	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	549.611	0.000	...	
3	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	623.362	6.054	...	
5	14.23	15.350	85.518	1171.604	198.538	344.014	325.195	1.436	628.245	6.020	...	
6	13.49	13.700	98.186	1243.688	116.275	346.208	326.982	1.434	696.766	0.000	...	
...	...	...	...	...	...	...	...	...	...	...	...	
276	22.70	15.517	83.008	1288.010	306.886	350.155	322.485	1.590	568.752	6.170	...	
296	20.50	13.358	97.662	1304.597	377.678	347.672	313.147	1.546	496.460	6.340	...	
297	20.40	14.233	89.790	1278.006	379.458	354.290	315.558	1.515	491.374	0.000	...	
298	20.90	15.167	84.640	1283.706	339.440	354.803	311.041	1.635	532.419	6.340	...	
307	20.89	14.308	94.172	1327.832	251.120	351.263	332.485	1.522	631.514	0.000	...	

226 rows × 22 columns

In [ ]: