```
import pandas as pd
```

```
data = pd.read_excel('/content/01.Data Cleaning and Preprocessing.xlsx')
```

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 324 entries, 0 to 323
Data columns (total 23 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Observation     324 non-null    object 
 1   Y-Kappa         324 non-null    float64
 2   ChipRate        319 non-null    float64
 3   BF-CMratio      307 non-null    float64
 4   BlowFlow        308 non-null    float64
 5   ChipLevel4      323 non-null    float64
 6   T-upperExt-2    322 non-null    float64
 7   T-lowerExt-2    322 non-null    float64
 8   UCZAA           299 non-null    float64
 9   WhiteFlow-4     323 non-null    float64
 10  AAWhiteSt-4     173 non-null    float64
 11  AA-Wood-4       323 non-null    float64
 12  ChipMoisture-4  323 non-null    float64
 13  SteamFlow-4     323 non-null    float64
 14  Lower-HeatT-3   322 non-null    float64
 15  Upper-HeatT-3   322 non-null    float64
 16  ChipMass-4      323 non-null    float64
 17  WeakLiquorF     323 non-null    float64
 18  BlackFlow-2     322 non-null    float64
 19  WeakWashF       323 non-null    float64
 20  SteamHeatF-3    322 non-null    float64
 21  T-Top-Chips-4   323 non-null    float64
 22  SulphidityL-4   173 non-null    float64
dtypes: float64(22), object(1)
memory usage: 58.3+ KB
```

```
data.describe()
```

| | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | AAWhiteSt-4 | ... | SteamFl |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **count** | 324.000000 | 319.000000 | 307.000000 | 308.000000 | 323.000000 | 322.000000 | 322.000000 | 299.000000 | 323.000000 | 173.000000 | ... | 323.000 |
| **mean** | 20.635370 | 14.347937 | 87.464456 | 1237.837614 | 258.164483 | 356.904295 | 324.020180 | 1.492010 | 591.732260 | 6.140410 | ... | 66.668 |
| **std** | 3.070036 | 1.499095 | 7.995012 | 100.593735 | 87.987452 | 9.209290 | 7.621402 | 0.105923 | 67.016351 | 0.081609 | ... | 5.708 |
| **min** | 12.170000 | 9.983000 | 68.645000 | 0.000000 | 0.000000 | 339.168000 | 284.633000 | 1.182000 | 405.111000 | 5.890000 | ... | 48.568 |
| **25%** | 18.382500 | 13.358000 | 81.823000 | 1193.215250 | 213.527000 | 350.241250 | 321.420000 | 1.431500 | 540.989500 | 6.089000 | ... | 62.518 |
| **50%** | 20.845000 | 14.308000 | 86.739000 | 1273.138500 | 271.792000 | 356.843000 | 325.669000 | 1.498000 | 592.895000 | 6.135000 | ... | 67.429 |
| **75%** | 23.032500 | 15.517000 | 92.372000 | 1289.196000 | 321.680000 | 362.242250 | 329.175000 | 1.560500 | 639.480500 | 6.199000 | ... | 71.522 |
| **max** | 27.600000 | 16.958000 | 121.717000 | 1351.240000 | 419.014000 | 399.135000 | 337.012000 | 1.747000 | 731.394000 | 6.340000 | ... | 76.147 |

8 rows × 22 columns

```
data.notnull().sum()
```

```
Observation       324
Y-Kappa           324
ChipRate          319
BF-CMratio        307
BlowFlow          308
ChipLevel4        323
T-upperExt-2      322
T-lowerExt-2      322
UCZAA             299
WhiteFlow-4       323
AAWhiteSt-4       173
AA-Wood-4         323
ChipMoisture-4    323
SteamFlow-4       323
Lower-HeatT-3     322
Upper-HeatT-3     322
```

```
ChipMass-4        323
WeakLiquorF       323
BlackFlow-2       322
WeakWashF         323
SteamHeatF-3      322
T-Top-Chips-4     323
SulphidityL-4     173
dtype: int64
```

```
data.isnull()
```

| | Observation | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | ... | SteamFlow-4 | Lower-HeatT-3 | Upper HeatT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **1** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **2** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **3** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **4** | False | False | False | False | False | False | False | False | True | False | ... | False | False | Fals |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | . |
| **319** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **320** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **321** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **322** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |
| **323** | False | False | False | False | False | False | False | False | False | False | ... | False | False | Fals |

324 rows × 23 columns

```
data.isnull().sum()
```

```
Observation         0
Y-Kappa             0
ChipRate            5
BF-CMratio         17
BlowFlow           16
ChipLevel4          1
T-upperExt-2        2
T-lowerExt-2        2
UCZAA              25
WhiteFlow-4         1
AAWhiteSt-4       151
AA-Wood-4           1
ChipMoisture-4      1
SteamFlow-4         1
Lower-HeatT-3       2
Upper-HeatT-3       2
ChipMass-4          1
WeakLiquorF         1
BlackFlow-2         2
WeakWashF           1
SteamHeatF-3        2
T-Top-Chips-4       1
SulphidityL-4     151
dtype: int64
```

```
data1 = data.fillna(value=0)
data1
```

| | Observation | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | ... | SteamFlow-4 | Lower-HeatT-3 | Uppe Heat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 31-00:00 | 23.10 | 16.520 | 121.717 | 1177.607 | 169.805 | 358.282 | 329.545 | 1.443 | 599.253 | ... | 67.122 | 329.432 | 303.0 |
| **1** | 31-01:00 | 27.60 | 16.810 | 79.022 | 1328.360 | 341.327 | 351.050 | 329.067 | 1.549 | 537.201 | ... | 60.012 | 330.823 | 304.8 |
| **2** | 31-02:00 | 23.19 | 16.709 | 79.562 | 1329.407 | 239.161 | 350.022 | 329.260 | 1.600 | 549.611 | ... | 61.304 | 329.140 | 303.3 |
| **3** | 31-03:00 | 23.60 | 16.478 | 81.011 | 1334.877 | 213.527 | 350.938 | 331.142 | 1.604 | 623.362 | ... | 68.496 | 328.875 | 302.2 |
| **4** | 31-04:00 | 22.90 | 15.618 | 93.244 | 1334.168 | 243.131 | 351.640 | 332.709 | 0.000 | 638.672 | ... | 70.022 | 328.352 | 300.9 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **319** | 10-16:00 | 23.75 | 12.667 | 93.450 | 1178.252 | 276.955 | 347.286 | 310.970 | 1.523 | 513.956 | ... | 61.141 | 330.117 | 304.0 |
| **320** | 9-19:00 | 19.80 | 12.558 | 94.352 | 1184.119 | 297.071 | 399.135 | 319.576 | 1.451 | 570.058 | ... | 67.667 | 330.848 | 304.6 |
| **321** | 9-20:00 | 23.01 | 12.550 | 90.842 | 1188.517 | 289.826 | 373.633 | 314.591 | 1.457 | 549.306 | ... | 66.446 | 330.226 | 304.6 |
| **322** | 9-21:00 | 24.32 | 13.083 | 88.910 | 1192.879 | 318.006 | 364.081 | 308.559 | 1.523 | 504.852 | ... | 61.054 | 327.346 | 304.3 |
| **323** | 9-22:00 | 25.75 | 13.417 | 85.451 | 1186.342 | 248.312 | 356.289 | 310.482 | 1.474 | 497.375 | ... | 58.247 | 328.092 | 304.0 |

324 rows × 23 columns

```
data1.isnull().sum()
```

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

```
data2 = data.fillna(method = 'bfill')
data2
```

| | Observation | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | ... | SteamFlow-4 | Lower-HeatT-3 | Uppe Heat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 31-00:00 | 23.10 | 16.520 | 121.717 | 1177.607 | 169.805 | 358.282 | 329.545 | 1.443 | 599.253 | ... | 67.122 | 329.432 | 303.0 |
| **1** | 31-01:00 | 27.60 | 16.810 | 79.022 | 1328.360 | 341.327 | 351.050 | 329.067 | 1.549 | 537.201 | ... | 60.012 | 330.823 | 304.8 |
| **2** | 31-02:00 | 23.19 | 16.709 | 79.562 | 1329.407 | 239.161 | 350.022 | 329.260 | 1.600 | 549.611 | ... | 61.304 | 329.140 | 303.3 |
| **3** | 31-03:00 | 23.60 | 16.478 | 81.011 | 1334.877 | 213.527 | 350.938 | 331.142 | 1.604 | 623.362 | ... | 68.496 | 328.875 | 302.2 |
| **4** | 31-04:00 | 22.90 | 15.618 | 93.244 | 1334.168 | 243.131 | 351.640 | 332.709 | 1.436 | 638.672 | ... | 70.022 | 328.352 | 300.9 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **319** | 10-16:00 | 23.75 | 12.667 | 93.450 | 1178.252 | 276.955 | 347.286 | 310.970 | 1.523 | 513.956 | ... | 61.141 | 330.117 | 304.0 |
| **320** | 9-19:00 | 19.80 | 12.558 | 94.352 | 1184.119 | 297.071 | 399.135 | 319.576 | 1.451 | 570.058 | ... | 67.667 | 330.848 | 304.6 |
| **321** | 9-20:00 | 23.01 | 12.550 | 90.842 | 1188.517 | 289.826 | 373.633 | 314.591 | 1.457 | 549.306 | ... | 66.446 | 330.226 | 304.6 |
| **322** | 9-21:00 | 24.32 | 13.083 | 88.910 | 1192.879 | 318.006 | 364.081 | 308.559 | 1.523 | 504.852 | ... | 61.054 | 327.346 | 304.3 |
| **323** | 9-22:00 | 25.75 | 13.417 | 85.451 | 1186.342 | 248.312 | 356.289 | 310.482 | 1.474 | 497.375 | ... | 58.247 | 328.092 | 304.0 |

324 rows × 23 columns

```
data3 = data.fillna(method = 'pad')
data3
```

| | Observation | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | ... | SteamFlow-4 | Lower-HeatT-3 | Uppe Heat |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 31-00:00 | 23.10 | 16.520 | 121.717 | 1177.607 | 169.805 | 358.282 | 329.545 | 1.443 | 599.253 | ... | 67.122 | 329.432 | 303.0 |
| **1** | 31-01:00 | 27.60 | 16.810 | 79.022 | 1328.360 | 341.327 | 351.050 | 329.067 | 1.549 | 537.201 | ... | 60.012 | 330.823 | 304.8 |
| **2** | 31-02:00 | 23.19 | 16.709 | 79.562 | 1329.407 | 239.161 | 350.022 | 329.260 | 1.600 | 549.611 | ... | 61.304 | 329.140 | 303.3 |
| **3** | 31-03:00 | 23.60 | 16.478 | 81.011 | 1334.877 | 213.527 | 350.938 | 331.142 | 1.604 | 623.362 | ... | 68.496 | 328.875 | 302.2 |
| **4** | 31-04:00 | 22.90 | 15.618 | 93.244 | 1334.168 | 243.131 | 351.640 | 332.709 | 1.604 | 638.672 | ... | 70.022 | 328.352 | 300.9 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **319** | 10-16:00 | 23.75 | 12.667 | 93.450 | 1178.252 | 276.955 | 347.286 | 310.970 | 1.523 | 513.956 | ... | 61.141 | 330.117 | 304.0 |
| **320** | 9-19:00 | 19.80 | 12.558 | 94.352 | 1184.119 | 297.071 | 399.135 | 319.576 | 1.451 | 570.058 | ... | 67.667 | 330.848 | 304.6 |
| **321** | 9-20:00 | 23.01 | 12.550 | 90.842 | 1188.517 | 289.826 | 373.633 | 314.591 | 1.457 | 549.306 | ... | 66.446 | 330.226 | 304.6 |
| **322** | 9-21:00 | 24.32 | 13.083 | 88.910 | 1192.879 | 318.006 | 364.081 | 308.559 | 1.523 | 504.852 | ... | 61.054 | 327.346 | 304.3 |
| **323** | 9-22:00 | 25.75 | 13.417 | 85.451 | 1186.342 | 248.312 | 356.289 | 310.482 | 1.474 | 497.375 | ... | 58.247 | 328.092 | 304.0 |

324 rows × 23 columns

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
```

```
data1.drop(['Observation'],axis = 1, inplace = True)
data1.columns
```

```
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')
```

```
Q1 = data1.quantile(0.25)
Q3 =  data1.quantile(0.75)
Q = Q3 - Q1
Q
```

```
Y-Kappa             4.65000
ChipRate            2.25625
```

```
BF-CMratio         11.11225
BlowFlow           98.43375
ChipLevel4        107.92275
T-upperExt-2       11.96500
T-lowerExt-2        7.82875
UCZAA               0.13925
WhiteFlow-4        98.59525
AAWhiteSt-4         6.14000
AA-Wood-4           1.45900
ChipMoisture-4      2.22000
SteamFlow-4         9.04675
Lower-HeatT-3       8.46750
Upper-HeatT-3       7.77050
ChipMass-4         19.70375
WeakLiquorF       174.05550
BlackFlow-2       276.51675
WeakWashF         271.44325
SteamHeatF-3        6.94975
T-Top-Chips-4       2.01025
SulphidityL-4      30.40250
dtype: float64
```

```
data1 = data1[~((data1 < Q1 - 1.5 * Q) | (data1 > Q3 + 1.5 * Q)).any(axis=1)]
```

```
data1
```

| | Y-Kappa | ChipRate | BF-CMratio | BlowFlow | ChipLevel4 | T-upperExt-2 | T-lowerExt-2 | UCZAA | WhiteFlow-4 | AAWhiteSt-4 | ... | SteamFlow-4 | Lower-HeatT-3 | Upper HeatT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 27.60 | 16.810 | 79.022 | 1328.360 | 341.327 | 351.050 | 329.067 | 1.549 | 537.201 | 6.076 | ... | 60.012 | 330.823 | 304.87 |
| **2** | 23.19 | 16.709 | 79.562 | 1329.407 | 239.161 | 350.022 | 329.260 | 1.600 | 549.611 | 0.000 | ... | 61.304 | 329.140 | 303.38 |
| **3** | 23.60 | 16.478 | 81.011 | 1334.877 | 213.527 | 350.938 | 331.142 | 1.604 | 623.362 | 6.054 | ... | 68.496 | 328.875 | 302.25 |
| **5** | 14.23 | 15.350 | 85.518 | 1171.604 | 198.538 | 344.014 | 325.195 | 1.436 | 628.245 | 6.020 | ... | 65.225 | 322.103 | 298.51 |