

```
In [ ]: import pandas as pd

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

<>::1: SyntaxWarning: invalid escape sequence '\h'
<<::1: SyntaxWarning: invalid escape sequence '\h'
C:\Users\hp\AppData\Local\Temp\ipykernel_18412\57711915.py:1: SyntaxWarning: invalid escape sequence '\h'
data = pd.read_csv('C:\\Users\\hp\\Downloads\\01.Data Cleaning and Preprocessing.csv')

Out [ ]:
      Observation  Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  T-lowerExt-2  UCZAA  WhiteFlow-4  ...  SteamFlow-4  Lower-HeatT-3  Upper-HeatT-3  ChipMass-4  WeakLiquorF  BlackFlow-2  WeakWashF  SteamHeatF-3  T-Top-Chips-4  SulphidityL-4
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    303.099    175.964    1127.197    1319.039    257.325    54.612    252.077    NaN
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    304.879    163.202    665.975    1297.317    241.182    46.603    251.406    29.11
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    303.383    164.013    677.534    1327.072    237.272    51.795    251.335    NaN
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    302.254    181.487    767.853    1324.461    239.478    54.846    250.312    29.02
4      31-04.00   22.90    15.618    93.244   1334.168    243.131    351.640    332.709    NaN    638.672  ...      70.022    328.352    300.954    183.929    888.448    1343.424    215.372    54.186    249.916    29.01
...
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    304.006    148.174    1027.201    1317.271    381.643    45.264    252.947    30.86
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    304.616    165.178    906.962    1311.177    25.494    50.528    252.092    30.70
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    304.686    160.841    887.125    1319.226    0.638    45.549    252.438    NaN
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    304.363    147.589    804.423    1320.225    0.000    43.725    253.176    31.13
323    9-22.00   25.75    13.417    85.451   1184.342    248.312    356.289    310.482    1.474    497.375  ...      58.247    328.092    304.093    144.218    828.328    1320.848    1.276    43.840    253.216    NaN

324 rows x 23 columns
```

```
In [ ]: #Info function gives both data-type and total of non-null values of each column
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  AWhiteSt-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: 58.5+ KB

In [ ]: #To print all the columns name
data.columns

Out [ ]: Index(['Observation', 'Y-Kappa', 'ChipRate', 'BF-CMratio', 'BlowFlow', 'ChipLevel4', 'T-upperExt-2', 'T-lowerExt-2', 'UCZAA', 'WhiteFlow-4', 'AWhiteSt-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'],
      dtypes=object)
```

```
In [ ]: #Deleting the column
data.drop(['Observation'], axis=1, inplace = True )

In [ ]: #Observation column is deleted
data.columns

Out [ ]: Index(['Y-Kappa', 'ChipRate', 'BF-CMratio', 'BlowFlow', 'ChipLevel4', 'T-upperExt-2', 'T-lowerExt-2', 'UCZAA', 'WhiteFlow-4', 'AWhiteSt-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'],
      dtypes=object)
```

```
In [ ]: #Total number of null values in each column
data.isnull().sum()

Out [ ]:
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
AWhiteSt-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

In [ ]: #Total number of null values in dataset
data.isnull().sum().sum()

Out [ ]: 388

In [ ]: #Dropping all duplicates values in dataset
data = data.drop_duplicates()
data

Out [ ]:
      Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  T-lowerExt-2  UCZAA  WhiteFlow-4  AWhiteSt-4  ...  SteamFlow-4  Lower-HeatT-3  Upper-HeatT-3  ChipMass-4  WeakLiquorF  BlackFlow-2  WeakWashF  SteamHeatF-3  T-Top-Chips-4  SulphidityL-4
0      23.10    16.520    121.717   1177.607    169.805    358.282    329.545    1.443    599.253    NaN  ...      67.122    329.432    303.099    175.964    1127.197    1319.039    257.325    54.612    252.077    NaN
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    304.879    163.202    665.975    1297.317    241.182    46.603    251.406    29.11
2      23.19    16.709    79.562   1329.407    239.161    350.022    329.260    1.600    549.611    NaN  ...      61.304    329.140    303.383    164.013    677.534    1327.072    237.272    51.795    251.335    NaN
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    302.254    181.487    767.853    1324.461    239.478    54.846    250.312    29.02
4      22.90    15.618    93.244   1334.168    243.131    351.640    332.709    NaN    638.672    6.110  ...      70.022    328.352    300.954    183.929    888.448    1343.424    215.372    54.186    249.916    29.01
...
298    20.90    15.167    84.640   1283.706    339.440    354.803    311.041    1.635    532.419    6.340  ...      65.561    332.924    307.626    145.299    832.906    1344.708    388.911    49.524    251.833    30.29
299    24.98    NaN      85.034   1278.345    368.564    357.723    321.387    NaN    520.365    6.220  ...      65.729    332.523    307.169    151.544    905.639    1344.469    418.979    48.135    251.614    30.47
300    21.00    NaN      88.013   1307.722    278.842    357.438    323.757    NaN    553.070    6.230  ...      65.795    331.263    306.400    157.954    908.691    1344.588    462.712    54.373    251.197    NaN
301    21.40    NaN      85.490   1255.986    273.484    361.365    322.689    NaN    590.199    6.230  ...      71.456    333.032    308.732    174.069    986.206    1348.747    457.313    53.194    251.324    30.46
307    20.89    14.308    94.172   1327.832    251.120    351.263    332.485    1.522    631.514    NaN  ...      71.286    328.699    300.706    180.229    903.605    1323.082    232.729    54.503    250.084    NaN

301 rows x 22 columns
```

```
In [ ]: #Detecting and removing the outliers in dataset
q1 = data.quantile(0.25)
q3 = data.quantile(0.75)
iqr = q3 - q1

lower_bound = q1 - (1.5*iqr)
upper_bound = q3 + (1.5*iqr)

outliers = data[(data < lower_bound)|(data > upper_bound)] #outliers

data_no_outliers = data[(data > lower_bound)&(data < upper_bound)] #dataset with no outliers
print(data_no_outliers)

      Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  \
0      23.10    16.520      NaN    1177.607    169.805    358.282
1      27.60    16.810    79.022   1328.360    341.327    351.050
2      23.19    16.709    79.562   1329.407    239.161    350.022
3      23.60    16.478    81.011   1334.877    213.527    350.938
4      22.90    15.618    93.244   1334.168    243.131    351.640
...
298    20.90    15.167    84.640   1283.706    339.440    354.803
299    24.98    NaN      85.034   1278.345    368.564    357.723
300    21.00    NaN      88.013   1307.722    278.842    357.438
301    21.40    NaN      85.490   1255.986    273.484    361.365
307    20.89    14.308    94.172   1327.832    251.120    351.263

      T-lowerExt-2  UCZAA  WhiteFlow-4  AWhiteSt-4  ...  SteamFlow-4  \
0      329.545    1.443    599.253      NaN  ...      67.122
1      329.067    1.549    537.201    6.076  ...      60.012
2      329.260    1.600    549.611      NaN  ...      61.304
3      331.142    1.604    623.362    6.054  ...      68.496
4      332.709      NaN    638.672    6.110  ...      70.022
...
298    311.041    1.635    532.419    6.340  ...      65.561
299    321.387      NaN    520.365    6.220  ...      65.729
300    323.757      NaN    553.070      NaN  ...      65.795
301    322.689      NaN    590.199    6.230  ...      71.456
307    332.485    1.522    631.514      NaN  ...      71.286

      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
...
298    332.924    307.626    145.299    832.906    1344.708
299    332.523    307.169    151.544    905.639    1344.469
300    331.263    306.400    157.954    908.691    1344.588
301    333.032    308.732    174.069    986.206    1348.747
307    328.699    300.706    180.229    903.605    1323.082

      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      NaN
...
298    388.911    49.524    251.833    30.29
299    418.979    48.135    251.614    30.47
300    462.712    54.373    251.197      NaN
301    457.313    53.194    251.324    30.46
307    232.729    54.503    250.084      NaN

[301 rows x 22 columns]
```

```
In [ ]: #Dropping all the rows that with null values
data_no_outliers.dropna()

Out [ ]:
      Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  T-lowerExt-2  UCZAA  WhiteFlow-4  AWhiteSt-4  ...  SteamFlow-4  Lower-HeatT-3  Upper-HeatT-3  ChipMass-4  WeakLiquorF  BlackFlow-2  WeakWashF  SteamHeatF-3  T-Top-Chips-4  SulphidityL-4
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    304.879    163.202    665.975    1297.317    241.182    46.603    251.406    29.11
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    302.254    181.487    767.853    1324.461    239.478    54.846    250.312    29.02
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    298.517    165.814    826.243    907.641    595.875    52.807    249.580    30.34
7      22.65    14.100    91.887    1307.852    288.989    352.321    331.162    1.468    625.549    6.143  ...      71.298    329.662    301.539    179.886    837.178    1315.111    234.047    53.805    249.971    29.22
9      24.70    13.850    96.208    1334.892    362.511    352.372    327.358    1.515    553.172    6.199  ...      64.249    332.264    305.419    166.120    908.810    1318.725    180.375    48.842    251.121    29.21
...
270    21.25    14.192    91.907    1286.693    346.418    349.544    318.284    1.598    551.442    6.180  ...      68.053    332.175    308.950    158.412    1038.187    1366.592    278.115    50.512    252.504    30.47
274    23.17    15.542    82.392    1280.501    334.242    348.771    316.870    1.596    558.715    6.220  ...      68.318    331.510    309.023    156.631    855.558    1367.960    300.409    49.837    252.626    30.48
276    22.70    15.517    83.008    1288.010    306.886    350.155    322.485    1.590    568.752    6.170  ...      67.678    331.854    309.346    160.061    910.013    1381.389    441.934    51.466    252.216    29.59
296    20.50    13.358    97.662    1304.597    377.678    347.672    313.147    1.546    496.460    6.340  ...      60.119    332.615    308.575    141.076    997.904    1334.703    389.497    46.206    252.423    30.43
298    20.90    15.167    84.640    1283.706    339.440    354.803    311.041    1.635    532.419    6.340  ...      65.561    332.924    307.626    145.299    832.906    1344.708    388.911    49.524    251.833    30.29

107 rows x 22 columns
```

```
In [ ]: #Replacing the null values by its following value or by its preceding value
data_no_outliers = data_no_outliers.bfill() #backwardfill
data_no_outliers = data_no_outliers.ffill() #forwardfill
data_no_outliers

Out [ ]:
      Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  T-lowerExt-2  UCZAA  WhiteFlow-4  AWhiteSt-4  ...  SteamFlow-4  Lower-HeatT-3  Upper-HeatT-3  ChipMass-4  WeakLiquorF  BlackFlow-2  WeakWashF  SteamHeatF-3  T-Top-Chips-4  SulphidityL-4
0      23.10    16.520    79.022   1177.607    169.805    358.282    329.545    1.443    599.253    6.076  ...      67.122    329.432    303.099    175.964    1127.197    1319.039    257.325    54.612    252.077    29.11
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    304.879    163.202    665.975    1297.317    241.182    46.603    251.406    29.11
2      23.19    16.709    79.562   1329.407    239.161    350.022    329.260    1.600    549.611    6.054  ...      61.304    329.140    303.383    164.013    677.534    1327.072    237.272    51.795    251.335    29.02
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    302.254    181.487    767.853    1324.461    239.478    54.846    250.312    29.02
4      22.90    15.618    93.244   1334.168    243.131    351.640    332.709    1.436    638.672    6.110  ...      70.022    328.352    300.954    183.929    888.448    1343.424    215.372    54.186    249.916    30.34
...
298    20.90    15.167    84.640   1283.706    339.440    354.803    311.041    1.635    532.419    6.340  ...      65.561    332.924    307.626    145.299    832.906    1344.708    388.911    49.524    251.833    30.29
299    24.98    14.308    85.034   1278.345    368.564    357.723    321.387    1.522    520.365    6.220  ...      65.729    332.523    307.169    151.544    905.639    1344.469    418.979    48.135    251.614    30.47
300    21.00    14.308    88.013   1307.722    278.842    357.438    323.757    1.522    553.070    6.230  ...      65.795    331.263    306.400    157.954    908.691    1344.588    462.712    54.373    251.197    30.46
301    21.40    14.308    85.490   1255.986    273.484    361.365    322.689    1.522    590.199    6.230  ...      71.456    333.032    308.732    174.069    986.206    1348.747    457.313    53.194    251.324    30.46
307    20.89    14.308    94.172   1327.832    251.120    351.263    332.485    1.522    631.514    6.230  ...      71.286    328.699    300.706    180.229    903.605    1323.082    232.729    54.503    250.084    30.46

301 rows x 22 columns
```

```
In [ ]: #Checking is there any null values in the dataset.
data_no_outliers.isnull().sum()

Out [ ]: 0

In [ ]: #Describe function gives the summary statistics of dataset
data_no_outliers.describe()

Out [ ]:
      Y-Kappa  ChipRate  BF-CMratio  BlowFlow  ChipLevel4  T-upperExt-2  T-lowerExt-2  UCZAA  WhiteFlow-4  AWhiteSt-4  ...  SteamFlow-4  Lower-HeatT-3  Upper-HeatT-3  ChipMass-4  WeakLiquorF  BlackFlow-2  WeakWashF  SteamHeatF-3  T-Top-Chips-4  SulphidityL-4
count  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  ...  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000  301.000000
mean    20.568005  14.337565  86.937296  1240.418412  268.989761  356.718465  324.556246  1.493272  593.044704  6.146346  ...  66.857877  325.381272  300.386043  162.29295  876.333801  1170.471375  266.819492  49.699302  251.220252  30.453047
std     2.990751  1.473267  6.815905  61.889764  74.818251  8.313861  6.171239  0.101936  66.862119  0.078591  ...  5.546581  4.633449  4.640049  13.52897  120.403619  149.047109  163.422026  4.536224  1.301328  0.622187
min     12.170000  10.625000  68.645000  1078.498000  61.783000  339.168000  310.421000  1.234000  405.111000  5.940000  ...  50.663000  318.051000  293.312000  124.68200  537.597000  838.948000  0.000000  38.283000  248.359000  29.020000
25%    18.450000  13.358000  81.824000  1193.646000  221.306000  350.317000  321.539000  1.431000  543.235000  6.095000  ...  62.860000  321.235000  306.435000  152.91800  791.758000  1038.527000  142.5
```