```
#importing pandas library
import pandas as pd


#reading the csv file
data=pd.read_csv("/content/01.Data Cleaning and Preprocessing (2).csv")


#displaying the first five rows
data.head()
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UC
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	I

5 rows × 23 columns

```
#displaying the last five rows
data.tail()
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482

5 rows × 23 columns

```
#displaying information about the dataset
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
```

```
#displaying the datatype of columns
data.dtypes
```

```
Observation         object
Y-Kappa            float64
ChipRate           float64
BF-CMratio         float64
BlowFlow           float64
ChipLevel4         float64
T-upperExt-2       float64
T-lowerExt-2       float64
UCZAA              float64
WhiteFlow-4        float64
AAWhiteSt-4        float64
AA-Wood-4          float64
ChipMoisture-4     float64
SteamFlow-4        float64
Lower-HeatT-3      float64
Upper-HeatT-3      float64
ChipMass-4         float64
WeakLiquorF        float64
BlackFlow-2        float64
WeakWashF          float64
SteamHeatF-3       float64
T-Top-Chips-4      float64
```

```
SulphidityL-4      float64
dtype: object
```

```
#displaying the number of rows and columns
data.shape
```

```
(324, 23)
```

```
#column names
data.columns
```

```
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')
```

```
#descriptive statistics
data.describe()
```

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
count	324.000000	319.000000	307.000000	308.000000	323.000000	322.000000	322.000000
mean	20.635370	14.347937	87.464456	1237.837614	258.164483	356.904295	324.020180
std	3.070036	1.499095	7.995012	100.593735	87.987452	9.209290	7.621402
min	12.170000	9.983000	68.645000	0.000000	0.000000	339.168000	284.633000
25%	18.382500	13.358000	81.823000	1193.215250	213.527000	350.241250	321.420000
50%	20.845000	14.308000	86.739000	1273.138500	271.792000	356.843000	325.669000
75%	23.032500	15.517000	92.372000	1289.196000	321.680000	362.242250	329.175000
max	27.600000	16.958000	121.717000	1351.240000	419.014000	399.135000	337.012000

8 rows × 22 columns

```
#checking for duplicate rows
data.duplicated()
```

```
0      False
1      False
2      False
3      False
4      False
```

```
...
319     True
320     True
321     True
322     True
323     True
Length: 324, dtype: bool
```

```
#total number of duplicates
data.duplicated().sum()
```

```
23
```

```
#removing the duplicates
data.drop_duplicates()
```

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

301 rows × 23 columns

```
#check for null values
data.isnull()
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
319	False	False	False	False	False	False	False	False
320	False	False	False	False	False	False	False	False
321	False	False	False	False	False	False	False	False
322	False	False	False	False	False	False	False	False
323	False	False	False	False	False	False	False	False

324 rows × 23 columns

```
#null values in each column
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
```

```
#total number of null values in the dataset
data.isnull().sum().sum()
```

```
386
```

```
#check for non null values
data.notnull()
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	True	True	True	True	True	True	True	True
1	True	True	True	True	True	True	True	True
2	True	True	True	True	True	True	True	True
3	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True
...	...	...	...	...	...	...	...	...
319	True	True	True	True	True	True	True	True
320	True	True	True	True	True	True	True	True
321	True	True	True	True	True	True	True	True
322	True	True	True	True	True	True	True	True
323	True	True	True	True	True	True	True	True

324 rows × 23 columns

```
#non null values in each column
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
```

```
#replacing null values with 0
data1=data.fillna(value=0)
data1
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709
...	...	...	...	...	...	...	...	...
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482

324 rows × 23 columns

```
#forward fill null values
data2=data.fillna(method="pad")
data2
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709
...	...	...	...	...	...	...	...	...
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482

324 rows × 23 columns

```
#backward fill null values
data3=data.fillna(method="bfill")
data3
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709
...	...	...	...	...	...	...	...	...
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482

324 rows × 23 columns

```
#selecting numeric columns
data11=data.select_dtypes(include="number")
data11
```

	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteF
0	23.10	16.520	121.717	1177.607	169.805	358.282	329.545	1.443	59
1	27.60	16.810	79.022	1328.360	341.327	351.050	329.067	1.549	53
2	23.19	16.709	79.562	1329.407	239.161	350.022	329.260	1.600	54
3	23.60	16.478	81.011	1334.877	213.527	350.938	331.142	1.604	62
4	22.90	15.618	93.244	1334.168	243.131	351.640	332.709	NaN	63
...	...	...	...	...	...	...	...	...	
319	23.75	12.667	93.450	1178.252	276.955	347.286	310.970	1.523	51
320	19.80	12.558	94.352	1184.119	297.071	399.135	319.576	1.451	57
321	23.01	12.550	90.842	1188.517	289.826	373.633	314.591	1.457	54
322	24.32	13.083	88.910	1192.879	318.006	364.081	308.559	1.523	50
323	25.75	13.417	85.451	1186.342	248.312	356.289	310.482	1.474	49

324 rows × 22 columns

```
#replace null values with the mean of the column
data4=data.fillna(data11.mean())
data4
```

	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2
0	31-00:00	23.10	16.520	121.717	1177.607	169.805	358.282	329.545
1	31-01:00	27.60	16.810	79.022	1328.360	341.327	351.050	329.067
2	31-02:00	23.19	16.709	79.562	1329.407	239.161	350.022	329.260
3	31-03:00	23.60	16.478	81.011	1334.877	213.527	350.938	331.142
4	31-04:00	22.90	15.618	93.244	1334.168	243.131	351.640	332.709
...	...	...	...	...	...	...	...	...
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482

324 rows × 23 columns

```
#remove rows with any null values
data5=data.dropna()
data5
```

Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2

```
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
```

```
#removing observation column
data1.drop(['Observation'], axis=1, inplace=True)
```