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

In [2]:
```
df=pd.read_csv('01.Data Cleaning and Preprocessing.csv')
```

In [3]:
```
df
```

Out[3]:

	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	1357.271	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	1186.342	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 × 23 columns

Displaying first few rows

In [4]:
```
df.head()
```

Out[4]:

	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

5 rows × 23 columns

Filtering data based on condition

In [5]:
```
filtered_df = df[df['Y-Kappa'] > 25]
```

In [6]:
```
filtered_df
```

Out[6]:

	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
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	29.110
12	31-11:00	26.62	15.467	84.447	1334.255	386.971	349.392	321.021	1.428	531.250	...	59.407	330.284	303.248	156.797	799.947	1299.782	118.901	46.597	251.721	NaN
13	31-12:00	27.20	16.083	82.839	1332.331	366.855	350.094	327.439	1.486	527.893	...	60.271	330.023	302.883	160.562	771.158	1299.974	153.647	47.175	251.767	30.180
15	31-14:00	25.40	16.425	72.924	1197.775	118.821	350.765	329.799	1.635	585.011	...	65.474	329.773	302.884	175.646	756.154	1300.037	401.418	54.628	251.009	30.410
40	1-15:00	27.10	13.558	83.117	1175.417	289.256	339.168	318.386	1.360	480.184	...	48.568	318.228	294.850	131.537	744.659	996.046	118.899	41.985	253.450	NaN
64	2-15:00	25.40	16.958	74.405	1230.776	295.522	350.216	322.740	1.471	540.151	...	52.962	318.913	295.843	146.914	727.366	1052.879	476.150	43.701	252.097	NaN
80	3-07:00	26.50	16.300	75.411	1229.199	358.256	352.871	325.690	1.416	531.174	...	59.183	319.990	296.154	149.077	701.273	1026.333	259.773	42.596	250.775	31.910
121	5-00:00	25.56	14.900	84.953	1289.167	373.726	358.028	326.809	1.265	568.074	...	63.684	320.797	296.457	147.131	860.372	1014.646	149.973	43.781	252.652	31.796
123	5-02:00	25.30	15.225	82.030	1280.837	312.879	358.944	327.582	1.335	615.580	...	72.743	321.206	296.893	164.746	965.018	1043.596	213.414	49.987	251.192	30.595
151	6-06:00	26.02	15.800	79.293	1291.149	271.419	355.922	323.014	1.623	537.862	...	60.490	320.845	293.856	141.382	1099.121	1128.464	220.444	45.761	251.543	NaN
167	6-22:00	25.10	15.075	80.024	1281.051	343.214	354.905	325.323	1.377	618.276	...	59.783	320.077	295.551	157.611	901.062	1069.602	309.348	46.209	250.821	30.370
168	6-23:00	25.52	16.008	77.070	1285.782	338.656	355.149	325.669	1.462	619.954	...	63.048	320.280	295.959	158.476	878.072	1069.102	264.774	46.030	250.724	NaN
179	7-10:00	25.41	13.070	NaN	NaN	278.166	351.172	322.353	1.224	542.846	...	53.461	318.711	295.300	141.076	863.891	1091.979	43.288	48.221	252.042	30.180
193	8-00:00	26.10	15.083	78.098	1168.219	233.376	366.604	332.953	1.534	524.180	...	59.971	327.433	298.459	143.328	797.404	1176.439	497.742	46.377	250.287	NaN
239	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	304.093	144.218	828.328	1320.848	1.276	43.840	253.216	NaN
303	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	29.110
313	31-11:00	26.62	15.467	84.447	1334.255	386.971	349.392	321.021	1.428	531.250	...	59.407	330.284	303.248	156.797	799.947	1299.782	118.901	46.597	251.721	NaN
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	304.093	144.218	828.328	1320.848	1.276	43.840	253.216	NaN

18 rows × 23 columns

Handling missing values

In [7]:
```
df.isnull().sum()
```

Out[7]:
```
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
```

In [8]:
```
df['BF-CMratio'] = df['BF-CMratio'].fillna(df['BF-CMratio'].mean())
```

In [9]:
```
df.isnull().sum()
```

Out[9]:
```
Observation          0
Y-Kappa              0
ChipRate             5
BF-CMratio           0
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
```

In [10]:
```
df=df.dropna()
```

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

Out[11]:
```
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
```

Calculating summary statistics

In [12]:
```
df.describe()
```

Out[12]:

	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-3	ChipMass-4	WeakLiquorF	BlackFlow-2	WeakWashF	SteamHeatF-3	T-Top-Chips-4	Sulphi
count	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	...	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.000000	141.0
mean	20.780567	14.407277	86.842078	1241.878766	267.190312	356.841801	324.062270	1.497695	593.977879	6.144872	...	67.072582	325.484702	300.425092	162.551035	864.687546	1169.562624	271.164106	49.509894	251.274057	30.45
std	3.104667	1.501403	7.447316	72.414829	80.965627	9.924441	7.850917	0.103568	67.695750	0.080408	...	5.508386	4.586752	4.506755	14.200656	124.824316	151.937463	159.624109	4.648990	1.266735	0.68
min	12.170000	10.833000	68.645000	954.092000	0.000000	340.222000	287.705000	1.182000	405.111000	5.890000	...	51.192000	318.821000	293.312000	113.922000	493.856000	838.948000	0.000000	35.510000	248.390000	29.02
25%	18.500000	13.420000	81.087000	1194.181000	219.669000	350.155000	321.295000	1.437000	540.558000	6.093000	...	62.809000	321.448000	296.667000	153.147000	771.158000	1026.333000	149.973000	46.603000	250.312000	30.13
50%	20.750000	14.375000	86.950000	1276.082000	276.955000	357.367000	325.690000	1.500000	593.811000	6.140000	...	67.678000	323.430000	299.020000	163.258000	859.558000	1144.876000	282.044000	49.987000	251.406000	30.39
75%	23.300000	15.542000	92.142000	1290.006000	332.497000	362.372000	329.403000	1.556000	642.588000	6.199000	...	71.649000	329.851000	304.135000	173.204000	965.018000	1318.725000	392.864000	52.807000	252.423000	30.86
max	27.600000	16.817000	114.570000	1334.892000	419.014000	399.135000	333.706000	1.747000	731.394000	6.340000	...	76.147000	333.596000	310.487000	187.665000	1132.181000	1392.868000	598.171000	63.332000	254.122000	32.84

8 rows × 22 columns