

TASK-2

BY-PUSHP CHOUDHARY

Description:


This task involves using the Pandas library to manipulate data.

Responsibility:

Load a CSV file into a Pandas DataFrame. Perform operations like filtering data based on conditions, handling missing values, and calculating summary statistics.

```
import pandas as pd
```

```
data = pd.read_csv("//content//01.Data Cleaning and Preprocessing.csv") #read csv file
data
```



	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel4	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	..
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	
...
319	10-16:00	23.75	12.667	93.450	1178.252	276.955	347.286	310.970	1.523	513.956	
320	9-19:00	19.80	12.558	94.352	1184.119	297.071	399.135	319.576	1.451	570.058	
321	9-20:00	23.01	12.550	90.842	1188.517	289.826	373.633	314.591	1.457	549.306	
322	9-21:00	24.32	13.083	88.910	1192.879	318.006	364.081	308.559	1.523	504.852	
323	9-22:00	25.75	13.417	85.451	1186.342	248.312	356.289	310.482	1.474	497.375	

324 rows × 23 columns

```
type(data) #type of data
```

**pandas.core.frame.DataFrame**

```
def __init__(data=None, index: Axes | None=None, columns: Axes | None=None, dtype: Dtype | None=None, copy: bool | None=None) -> None
```

</usr/local/lib/python3.10/dist-packages/pandas/core/frame.py>

Two-dimensional, size-mutable, potentially heterogeneous tabular data.

Data structure also contains labeled axes (rows and columns).

Arithmetic operations align on both row and column labels. Can be thought of as a dict-like container for Series objects. The primary

```
data.info() #print the data's information
```



```
<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() #describe statistical
```



	Y-Kappa	ChipRate	BF- CMratio	BlowFlow	ChipLevel14	T- upperExt- 2	T- lowerExt- 2	UCZAA	WhiteFl
count	324.000000	319.000000	307.000000	308.000000	323.000000	322.000000	322.000000	299.000000	323.000
mean	20.635370	14.347937	87.464456	1237.837614	258.164483	356.904295	324.020180	1.492010	591.732
std	3.070036	1.499095	7.995012	100.593735	87.987452	9.209290	7.621402	0.105923	67.016
min	12.170000	9.983000	68.645000	0.000000	0.000000	339.168000	284.633000	1.182000	405.111
25%	18.382500	13.358000	81.823000	1193.215250	213.527000	350.241250	321.420000	1.431500	540.989
50%	20.845000	14.308000	86.739000	1273.138500	271.792000	356.843000	325.669000	1.498000	592.895
75%	23.032500	15.517000	92.372000	1289.196000	321.680000	362.242250	329.175000	1.560500	639.480
max	27.600000	16.958000	121.717000	1351.240000	419.014000	399.135000	337.012000	1.747000	731.394

8 rows × 22 columns

```
data =data.drop_duplicates() #drop all the duplicates
data
```



	Observation	Y- Kappa	ChipRate	BF- CMratio	BlowFlow	ChipLevel14	T- upperExt- 2	T- lowerExt- 2	UCZAA	WhiteFlow- 4	..
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

```
data.isnull() #true for null false for not null
```



	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel14	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	..
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

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



Observation	0
Y-Kappa	0
ChipRate	4
BF-CMratio	14
BlowFlow	13
ChipLevel14	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	

```
data.notnull() #true for not null and false for null
```



	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel14	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	..
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

```
data.isnull().sum().sum() #tells the number of null
```



352

```
data2 = data.fillna(value=0) #fill all the nulls to a value 0
data2
```



	Observation	Y-Kappa	ChipRate	BF-CMratio	BlowFlow	ChipLevel14	T-upperExt-2	T-lowerExt-2	UCZAA	WhiteFlow-4	..
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

```
data3 = data.fillna(method="pad") #forward filling
```

```
data4= data.fillna(method="bfill") #backward filling
```

```
import numpy as np
from scipy import stats
```

```
data2.columns #detect the outlier using IQR
```

```
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')
```

```
data2.drop(["Observation"],axis=1,inplace=True) #dropping unwanted column
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')
```

```
Q1 = data2.quantile(0.25)
Q3 = data2.quantile(0.75)
IQR = Q3 - Q1 #assigning value to IQR
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
```

```
data2 = data2[~((data2 < (Q1 - 1.5 * IQR)) |(data2 > (Q3 + 1.5 * IQR))).any(axis=1)] #formula for IQL
data2
```



	Y- Kappa	ChipRate	BF- CMratio	BlowFlow	ChipLevel14	T- upperExt- 2	T- lowerExt- 2	UCZAA	WhiteFlow- 4	AAWhiteSt- 4	...
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	..
...