

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

In [2]: `df=pd.read_csv('http://bit.ly/HDSC-Dataset')
df`

Out[2]:

	record_id	utility_id_ferc1	report_year	plant_name_ferc1	fuel_type_code_pudl	fuel_unit	fuel_qty_burned	fuel_mmbtu_per_unit	fuel_cost_
0	f1_fuel_1994_12_1_0_7	1	1994	rockport	coal	ton	5377489.0	16.590	
1	f1_fuel_1994_12_1_0_10	1	1994	rockport total plant	coal	ton	10486945.0	16.592	
2	f1_fuel_1994_12_2_0_1	2	1994	gorgas	coal	ton	2978683.0	24.130	
3	f1_fuel_1994_12_2_0_7	2	1994	barry	coal	ton	3739484.0	23.950	
4	f1_fuel_1994_12_2_0_10	2	1994	chickasaw	gas	mcf	40533.0	1.000	
...
29518	f1_fuel_2018_12_12_0_13	12	2018	neil simpson ct #1	gas	mcf	18799.0	1.059	
29519	f1_fuel_2018_12_12_1_1	12	2018	cheyenne prairie 58%	gas	mcf	806730.0	1.050	
29520	f1_fuel_2018_12_12_1_10	12	2018	lange ct facility	gas	mcf	104554.0	1.060	
29521	f1_fuel_2018_12_12_1_13	12	2018	wygen 3 bhp 52%	coal	ton	315945.0	16.108	
29522	f1_fuel_2018_12_12_1_14	12	2018	wygen 3 bhp 52%	gas	mcf	17853.0	1.059	

29523 rows × 11 columns

In [3]: `df.describe()`

Out[3]:

	utility_id_ferc1	report_year	fuel_qty_burned	fuel_mmbtu_per_unit	fuel_cost_per_unit_burned	fuel_cost_per_unit_delivered	fuel_cost_per_mmbtu
count	29523.000000	29523.000000	2.952300e+04	29523.000000	29523.000000	2.952300e+04	29523.000000
mean	118.601836	2005.806050	2.622119e+06	8.492111	208.649031	9.175704e+02	19.304354
std	74.178353	7.025483	9.118004e+06	10.600220	2854.490090	6.877593e+04	2091.540939
min	1.000000	1994.000000	1.000000e+00	0.000001	-276.080000	-8.749370e+02	-41.501000
25%	55.000000	2000.000000	1.381700e+04	1.024000	5.207000	3.778500e+00	1.940000
50%	122.000000	2006.000000	2.533220e+05	5.762694	26.000000	1.737100e+01	4.127000
75%	176.000000	2012.000000	1.424034e+06	17.006000	47.113000	4.213700e+01	7.745000
max	514.000000	2018.000000	5.558942e+08	341.260000	139358.000000	7.964521e+06	359278.000000

In [4]: `df.columns`

Out[4]: `Index(['record_id', 'utility_id_ferc1', 'report_year', 'plant_name_ferc1', 'fuel_type_code_pudl', 'fuel_unit', 'fuel_qty_burned', 'fuel_mmbtu_per_unit', 'fuel_cost_per_unit_burned', 'fuel_cost_per_unit_delivered', 'fuel_cost_per_mmbtu'], dtype='object')`

In [5]: `df.dtypes`

Out[5]:

record_id	object
utility_id_ferc1	int64
report_year	int64
plant_name_ferc1	object
fuel_type_code_pudl	object
fuel_unit	object
fuel_qty_burned	float64
fuel_mmbtu_per_unit	float64
fuel_cost_per_unit_burned	float64
fuel_cost_per_unit_delivered	float64
fuel_cost_per_mmbtu	float64
dtype:	object

In []: `Fuel_quantity_burned=
SkewValue= DataFrame.skew(axis=1)`

In [15]: `missing_values=df.isnull()
missing_values`

Out[15]:

	record_id	utility_id_ferc1	report_year	plant_name_ferc1	fuel_type_code_pudl	fuel_unit	fuel_qty_burned	fuel_mmbtu_per_unit	fuel_cost_per_unit_burne
0	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False
...
29518	False	False	False	False	False	False	False	False	False
29519	False	False	False	False	False	False	False	False	False
29520	False	False	False	False	False	False	False	False	False
29521	False	False	False	False	False	False	False	False	False
29522	False	False	False	False	False	False	False	False	False

29523 rows × 11 columns

In [17]: `missing_values=df.isnull().sum()
missing_values`

Out[17]:

record_id	0
utility_id_ferc1	0
report_year	0
plant_name_ferc1	0
fuel_type_code_pudl	0
fuel_unit	180
fuel_qty_burned	0
fuel_mmbtu_per_unit	0
fuel_cost_per_unit_burned	0
fuel_cost_per_unit_delivered	0
fuel_cost_per_mmbtu	0
dtype:	int64

In [18]: `df.groupby('fuel_unit')['fuel_unit'].count()`

Out[18]:

fuel_unit	
bbl	7998
gal	84
gramsU	464
kgU	110
mcf	11354
mmbtu	180
mdwth	95
mwhth	100
ton	8958
Name: fuel_unit, dtype: int64	

In []: