

Name - Tanishq Thuse

Year - SY

Branch - CSE(AI)

Div - B

Roll no. - 60

Submitted to : S M Jaybhaye Ma'am

## Assignment - 2

Problem Statement : Data Wrangling, IN Perform the following operations using Python on any open source dataset (e.g., data.csv)

- a. Import all the required Python Libraries.
- b. Locate open-source data from the web (e.g., <https://www.kaggle.com>). Provide a clear description of the data and its source (i.e., URL of the web site).
- c. Load the Dataset into pandas data frame.
- d. Data Preprocessing: check for missing values in the data using pandas `isnull()`, `describe()` function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.
- e. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.
- f. Turn categorical variables into quantitative variables in Python. Practical based on Data Loading, Storage and File Formats

### ✓ 1)Importing Libraries

```
# a) Import all the required Python Libraries.  
import pandas as pd  
import numpy as np  
from matplotlib import pyplot as plt
```

- 2) Locate open-source data from the web (e.g.,  
✓ <https://www.kaggle.com>). Provide a clear description of the data and its source (i.e., URL of the web site).

Dataset Link : <https://www.kaggle.com/datasets/joebeachcapital/carbon-majors-emissions-data>

The DataSet has 3 files : 1)emissions\_high\_granularity.csv 2)emissions\_low\_granularity.csv  
3)emissions\_medium\_granularity.csv

For this Assignment I am going to refer to 1)emissions\_high\_granularity.csv

```
file = "/content/emissions_high_granularity.csv"  
#I have to upload the file manually before running the cells
```

- ✓ 3) Load the Dataset into pandas data frame.

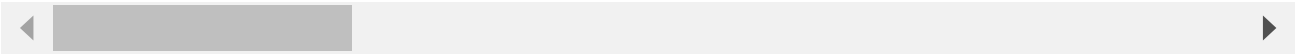
```
df = pd.read_csv(file)
```

- 4) Data Preprocessing: check for missing values in the data using  
✓ pandas isnull(), describe() function to get some initial statistics.  
Provide variable descriptions. Types of variables etc. Check the dimensions of the data frame.

```
df.head()
```



	year	parent_entity	parent_type	reporting_entity	commodity	production_value	proc
0	1962	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	0.9125	
1	1963	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	1.8250	
2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.3000	
3	1965	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	10.9500	
4	1966	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	13.5050	



Next steps:

Generate code with df

View recommended plots

New interactive sheet


df.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 15797 entries, 0 to 15796
Data columns (total 16 columns):
#   Column                                Non-Null Count  Dtype
---  -
```

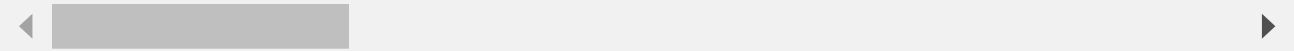
```
0 year 15797 non-null int64
1 parent_entity 15797 non-null object
2 parent_type 15797 non-null object
3 reporting_entity 15797 non-null object
4 commodity 15797 non-null object
5 production_value 15797 non-null float64
6 production_unit 15797 non-null object
7 product_emissions_MtCO2 15797 non-null float64
8 flaring_emissions_MtCO2 15797 non-null float64
9 venting_emissions_MtCO2 15797 non-null float64
10 own_fuel_use_emissions_MtCO2 15797 non-null float64
11 fugitive_methane_emissions_MtCO2e 15797 non-null float64
12 fugitive_methane_emissions_MtCH4 15797 non-null float64
13 total_operational_emissions_MtCO2e 15797 non-null float64
14 total_emissions_MtCO2e 15797 non-null float64
15 source 15797 non-null object
dtypes: float64(9), int64(1), object(6)
memory usage: 1.9+ MB
```

```
df.isnull()
```




	year	parent_entity	parent_type	reporting_entity	commodity	production_value
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
15792	False	False	False	False	False	False
15793	False	False	False	False	False	False
15794	False	False	False	False	False	False
15795	False	False	False	False	False	False
15796	False	False	False	False	False	False

15797 rows × 16 columns



```
df.isnull().sum()
```



```
year 0
parent_entity 0
parent_type 0
reporting_entity 0
commodity 0
production_value 0
production_unit 0
product_emissions_MtCO2 0
```

```

flaring_emissions_MtCO2      0
venting_emissions_MtCO2      0
own_fuel_use_emissions_MtCO2  0
fugitive_methane_emissions_MtCO2e  0
fugitive_methane_emissions_MtCH4  0
total_operational_emissions_MtCO2e  0
total_emissions_MtCO2e      0
source                        0
dtype: int64

```

```
df.describe()
```



	year	production_value	product_emissions_MtCO2	flaring_emissions_MtCO2
<b>count</b>	15797.000000	15797.000000	15797.000000	15797.000000
<b>mean</b>	1985.827942	327.879634	79.391514	0.517226
<b>std</b>	28.664256	1188.625001	261.984080	1.783744
<b>min</b>	1854.000000	0.000000	0.000000	0.000000
<b>25%</b>	1970.000000	11.800000	5.996490	0.000000
<b>50%</b>	1993.000000	59.970871	21.502409	0.015913
<b>75%</b>	2007.000000	246.375000	62.191954	0.197253
<b>max</b>	2022.000000	27192.000000	7769.222235	27.026872

```
df.dtypes
```



```

year                int64
parent_entity       object
parent_type         object
reporting_entity    object
commodity           object
production_value    float64
production_unit     object
product_emissions_MtCO2  float64
flaring_emissions_MtCO2  float64
venting_emissions_MtCO2  float64
own_fuel_use_emissions_MtCO2  float64
fugitive_methane_emissions_MtCO2e  float64
fugitive_methane_emissions_MtCH4  float64
total_operational_emissions_MtCO2e  float64
total_emissions_MtCO2e  float64
source             object
dtype: object

```

```
df.shape # to check dimensions of data frame
```



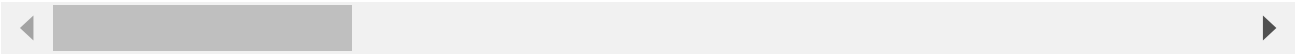
```
(15797, 16)
```

- 5) Data Formatting and Data Normalization: Summarize the types of variables by checking the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set. If variables are not in the correct data type, apply proper type conversions.

```
df.head()
```



	year	parent_entity	parent_type	reporting_entity	commodity	production_value	proc
0	1962	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	0.9125	
1	1963	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	1.8250	
2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.3000	
3	1965	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	10.9500	
4	1966	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	13.5050	



Next steps:

[Generate code with df](#)



[View recommended plots](#)

[New interactive sheet](#)

```
print(df.loc[0])
```



year	1962
parent_entity	Abu Dhabi National Oil Company
parent_type	State-owned Entity
reporting_entity	Abu Dhabi
commodity	Oil & NGL

```

production_value      0.9125
production_unit      Million bbl/yr
product_emissions_MtCO2  0.338928
flaring_emissions_MtCO2  0.005404
venting_emissions_MtCO2  0.001299
own_fuel_use_emissions_MtCO2  0.0
fugitive_methane_emissions_MtCO2e  0.018254
fugitive_methane_emissions_MtCH4  0.000652
total_operational_emissions_MtCO2e  0.024957
total_emissions_MtCO2e  0.363885
source                Abu Dhabi National Oil Company Annual Report 1...
Name: 0, dtype: object

```

```

print(df.iloc[ 0 : 1])
dataFrame = df

```

```

➡ year          parent_entity      parent_type reporting_entity \
0  1962  Abu Dhabi National Oil Company  State-owned Entity      Abu Dhabi

  commodity  production_value production_unit  product_emissions_MtCO2 \
0  Oil & NGL          0.9125  Million bbl/yr          0.338928

  flaring_emissions_MtCO2  venting_emissions_MtCO2 \
0          0.005404          0.001299

  own_fuel_use_emissions_MtCO2  fugitive_methane_emissions_MtCO2e \
0          0.0          0.018254

  fugitive_methane_emissions_MtCH4  total_operational_emissions_MtCO2e \
0          0.000652          0.024957

  total_emissions_MtCO2e          source
0          0.363885  Abu Dhabi National Oil Company Annual Report 1...

```

```
# print(df.loc['0'])
```

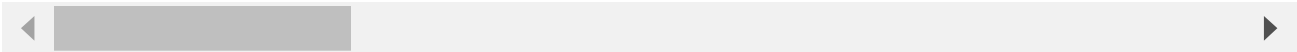
```
dataFrame.isnull()
```





	year	parent_entity	parent_type	reporting_entity	commodity	production_value
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
15792	False	False	False	False	False	False
15793	False	False	False	False	False	False
15794	False	False	False	False	False	False
15795	False	False	False	False	False	False
15796	False	False	False	False	False	False

15797 rows × 16 columns




```
dataFrame.dropna()
```



2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.30
3	1965	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	10.95
4	1966	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	13.50
...	...	...	...	...	...	
15792	2020	YPF	State-owned Entity	YPF	Natural Gas	394.00
15793	2021	YPF	State-owned Entity	YPF	Oil & NGL	90.00

15794	2021	YPF	State-owned Entity	YPF	Natural Gas	403.00
15795	2022	YPF	State-owned Entity	YPF	Oil & NGL	98.00
15796	2022	YPF	State-owned Entity	YPF	Natural Gas	423.00

```
print(dataFrame.loc[0])
```



year

parent\_entity

parent\_type

reporting\_entity

commodity

production\_value

production\_unit

product\_emissions\_MtCO2

flaring\_emissions\_MtCO2

venting\_emissions\_MtCO2

own\_fuel\_use\_emissions\_MtCO2

fugitive\_methane\_emissions\_MtCO2e

fugitive\_methane\_emissions\_MtCH4

total\_operational\_emissions\_MtCO2e

total\_emissions\_MtCO2e

source

Name: 0, dtype: object

1962

Abu Dhabi National Oil Company

State-owned Entity

Abu Dhabi

Oil & NGL

0.9125

Million bbl/yr

0.338928

0.005404

0.001299

0.0

0.018254

0.000652

0.024957

0.363885

Abu Dhabi National Oil Company Annual Report 1...

dataFrame



2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.30
3	1965	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	10.95
4	1966	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	13.50
...	...	...	...	...	...	
15792	2020	YPF	State-owned Entity	YPF	Natural Gas	394.00
15793	2021	YPF	State-owned Entity	YPF	Oil & NGL	90.00

15794	2021	YPF	State-owned Entity	YPF	Natural Gas	403.00
15795	2022	YPF	State-owned Entity	YPF	Oil & NGL	98.00
15796	2022	YPF	State-owned Entity	YPF	Natural Gas	423.00

Next steps:

Generate code with df

 View recommended plots

New interactive sheet

```
dates = pd.date_range("20130101", periods=6)
dates

↗ DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04',
                  '2013-01-05', '2013-01-06'],
                  dtype='datetime64[ns]', freq='D')

dataFrame
```



2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.30
3	1965	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	10.95
4	1966	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	13.50
...	...	...	...	...	...	
15792	2020	YPF	State-owned Entity	YPF	Natural Gas	394.00
15793	2021	YPF	State-owned Entity	YPF	Oil & NGL	90.00

15794	2021	YPF	State-owned Entity	YPF	Natural Gas	403.00
15795	2022	YPF	State-owned Entity	YPF	Oil & NGL	98.00
15796	2022	YPF	State-owned Entity	YPF	Natural Gas	423.00

Next steps:

[Generate code with df](#)

[View recommended plots](#)

[New interactive sheet](#)

dataFrame.dtypes

year	int64
parent_entity	object
parent_type	object
reporting_entity	object
commodity	object
production_value	float64
production_unit	object
product_emissions_MtCO2	float64
flaring_emissions_MtCO2	float64
venting_emissions_MtCO2	float64
own_fuel_use_emissions_MtCO2	float64
fugitive_methane_emissions_MtCO2e	float64
fugitive_methane_emissions_MtCH4	float64
total_operational_emissions_MtCO2e	float64
total_emissions_MtCO2e	float64
source	object
dtype:	object

df = dataFrame  
df.dtypes

year	int64
parent_entity	object
parent_type	object
reporting_entity	object
commodity	object
production_value	float64
production_unit	object
product_emissions_MtCO2	float64
flaring_emissions_MtCO2	float64

venting_emissions_MtCO2	float64
own_fuel_use_emissions_MtCO2	float64
fugitive_methane_emissions_MtCO2e	float64
fugitive_methane_emissions_MtCH4	float64
total_operational_emissions_MtCO2e	float64
total_emissions_MtCO2e	float64
source	object
dtype:	object

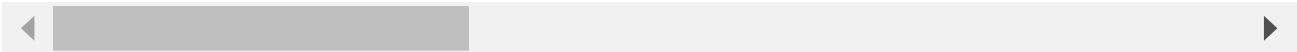
df.T





	0	1	2	3	4	
year	1962	1963	1964	1965	1966	
parent_entity	Abu Dhabi National Oil Company	Abu Dhabi National Oil Company	Abu Dhabi National Oil Company	Abu Dhabi National Oil Company	Abu Dhabi National Oil Company	[Na
parent_type	State-owned Entity	State-owned Entity	State-owned Entity	State-owned Entity	State-owned Entity	5 o' l
reporting_entity	Abu Dhabi	Abu Dhabi	Abu Dhabi	Abu Dhabi	Abu Dhabi	[
commodity	Oil & NGL	Oil & NGL	Oil & NGL	Oil & NGL	Oil & NGL	
production_value	0.9125	1.825	7.3	10.95	13.505	
production_unit	Million bbl/yr	Million bbl/yr	Million bbl/yr	Million bbl/yr	Million bbl/yr	M l
product_emissions_MtCO2	0.338928	0.677855	2.711422	4.067132	5.01613	5.42
flaring_emissions_MtCO2	0.005404	0.010808	0.043233	0.064849	0.07998	0.08
venting_emissions_MtCO2	0.001299	0.002598	0.010392	0.015588	0.019225	0.02
own_fuel_use_emissions_MtCO2	0.0	0.0	0.0	0.0	0.0	
fugitive_methane_emissions_MtCO2e	0.018254	0.036508	0.146033	0.219049	0.27016	0.29
fugitive_methane_emissions_MtCH4	0.000652	0.001304	0.005215	0.007823	0.009649	0.01
total_operational_emissions_MtCO2e	0.024957	0.049914	0.199657	0.299486	0.369366	0.39
total_emissions_MtCO2e	0.363885	0.72777	2.911079	4.366618	5.385495	5.82
source	Abu Dhabi National Oil Company Annual Report 1...	Abu Dhabi National Oil Company Annual Report 1...	Abu Dhabi National Oil Company Annual Report 1...	Abu Dhabi National Oil Company Annual Report 1...	Abu Dhabi National Oil Company Annual Report 1...	[Na Com Ai R

16 rows × 15797 columns



```
df.describe()
```



	year	production_value	product_emissions_MtCO2	flaring_emissions_MtCO2
count	15797.000000	15797.000000	15797.000000	15797.000000
mean	1985.827942	327.879634	79.391514	0.517226
std	28.664256	1188.625001	261.984080	1.783744
min	1854.000000	0.000000	0.000000	0.000000
25%	1970.000000	11.800000	5.996490	0.000000
50%	1993.000000	59.970871	21.502409	0.015913
75%	2007.000000	246.375000	62.191954	0.197253
max	2022.000000	27192.000000	7769.222235	27.026872

```
df.isnull()
```



	year	parent_entity	parent_type	reporting_entity	commodity	production_value
0	False	False	False	False	False	False
1	False	False	False	False	False	False
2	False	False	False	False	False	False
3	False	False	False	False	False	False
4	False	False	False	False	False	False
...	...	...	...	...	...	...
15792	False	False	False	False	False	False
15793	False	False	False	False	False	False
15794	False	False	False	False	False	False
15795	False	False	False	False	False	False
15796	False	False	False	False	False	False

15797 rows × 16 columns

Selection

```
#TO sort by values in product_emissions_MtCO2
df.sort_values(by='product_emissions_MtCO2', ascending=False)
```



	year	parent_entity	parent_type	reporting_entity	commodity	production_val
3662	2022	China (Coal)	Nation State	China (Coal)	Bituminous Coal	3185.8743
3658	2021	China (Coal)	Nation State	China (Coal)	Bituminous Coal	2882.6833
3654	2020	China (Coal)	Nation State	China (Coal)	Bituminous Coal	2725.8874
3650	2019	China (Coal)	Nation State	China (Coal)	Bituminous Coal	2669.8478
3626	2013	China (Coal)	Nation State	China (Coal)	Bituminous Coal	2620.0579
...	...	...	...	...	...	...
12862	1958	Saudi Aramco	State-owned Entity	Aramco	Natural Gas	0.0055
12860	1957	Saudi Aramco	State-owned Entity	Aramco	Natural Gas	0.0051
12858	1956	Saudi Aramco	State-owned Entity	Aramco	Natural Gas	0.0047
12856	1955	Saudi Aramco	State-owned Entity	Aramco	Natural Gas	0.0043
14190	1991	Suncor Energy	Investor-owned Company	Suncor Energy	Natural Gas	0.0000

```
df[0:3]
```



	year	parent_entity	parent_type	reporting_entity	commodity	production_value	proc
0	1962	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	0.9125	
1	1963	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	1.8250	
2	1964	Abu Dhabi National Oil Company	State-owned Entity	Abu Dhabi	Oil & NGL	7.3000	



Selection by label

```
df.loc[:, ['year','production_value' ]]
```