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: 9712 entries, 0 to 9711
```

```
Data columns (total 16 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   year                                  9712 non-null   int64
 1   parent_entity                         9712 non-null   object
 2   parent_type                           9712 non-null   object
 3   reporting_entity                      9712 non-null   object
 4   commodity                             9712 non-null   object
 5   production_value                      9712 non-null   float64
 6   production_unit                       9712 non-null   object
 7   product_emissions_MtCO2               9712 non-null   float64
 8   flaring_emissions_MtCO2               9712 non-null   float64
 9   venting_emissions_MtCO2               9712 non-null   float64
 10  own_fuel_use_emissions_MtCO2          9712 non-null   float64
 11  fugitive_methane_emissions_MtCO2e     9712 non-null   float64
 12  fugitive_methane_emissions_MtCH4      9712 non-null   float64
 13  total_operational_emissions_MtCO2e    9711 non-null   float64
 14  total_emissions_MtCO2e                9711 non-null   float64
 15  source                                9711 non-null   object
dtypes: float64(9), int64(1), object(6)
memory usage: 1.2+ 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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9707** | False | False | False | False | False | False |
| **9708** | False | False | False | False | False | False |
| **9709** | False | False | False | False | False | False |
| **9710** | False | False | False | False | False | False |
| **9711** | False | False | False | False | False | False |

9712 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  1
total_emissions_MtCO2e              1
source                              1
dtype: int64
```

```
df.describe()
```

| | year | production_value | product_emissions_MtCO2 | flaring_emissions_MtCO2 |
|---|---|---|---|---|
| **count** | 9712.000000 | 9712.000000 | 9712.000000 | 9712.000000 |
| **mean** | 1984.181322 | 360.334667 | 92.750838 | 0.504008 |
| **std** | 27.746984 | 1439.483609 | 317.321075 | 1.640646 |
| **min** | 1864.000000 | 0.007460 | 0.000399 | 0.000000 |
| **25%** | 1967.000000 | 14.383899 | 7.302732 | 0.000000 |
| **50%** | 1990.000000 | 66.557500 | 22.818983 | 0.018153 |
| **75%** | 2006.000000 | 262.262845 | 62.377203 | 0.189523 |
| **max** | 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
```

```
(9712, 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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9707** | False | False | False | False | False | False |
| **9708** | False | False | False | False | False | False |
| **9709** | False | False | False | False | False | False |
| **9710** | False | False | False | False | False | False |
| **9711** | False | False | False | False | False | False |

9712 rows × 16 columns

```
dataFrame.dropna()
```

| | year | parent_entity | parent_type | reporting_entity | commodity | production_value |
|---|---|---|---|---|---|---|
| **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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9706** | 2006 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 272.0000 |
| **9707** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 166.0000 |
| **9708** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 261.0000 |
| **9709** | 2008 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 193.8150 |
| | | | Investor- | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| **9710** | 2008 | Occidental Petroleum | owned Company | Occidental Petroleum | Natural Gas | 333.2450 |

9711 rows × 16 columns

```
print(dataFrame.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
```

```
dataFrame
```

| | year | parent_entity | parent_type | reporting_entity | commodity | production_value |
|---|---|---|---|---|---|---|
| **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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9707** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 166.0000 |
| **9708** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 261.0000 |
| **9709** | 2008 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 193.8150 |
| **9710** | 2008 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 333.2450 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **9711** | 2009 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 201.4800 |

9712 rows × 16 columns

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
```

| | year | parent_entity | parent_type | reporting_entity | commodity | production_value |
|---|---|---|---|---|---|---|
| **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 |
| **...** | ... | ... | ... | ... | ... | ... |
| **9707** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 166.0000 |
| **9708** | 2007 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 261.0000 |
| **9709** | 2008 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 193.8150 |
| **9710** | 2008 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Natural Gas | 333.2450 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **9711** | 2009 | Occidental Petroleum | Investor-owned Company | Occidental Petroleum | Oil & NGL | 201.4800 |

9712 rows × 16 columns

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 | D Na Com |
| **parent_type** | State-owned Entity | State-owned Entity | State-owned Entity | State-owned Entity | State-owned Entity | S o E |
| **reporting_entity** | Abu Dhabi | Abu Dhabi | Abu Dhabi | Abu Dhabi | Abu Dhabi | D |
| **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... | D Na Com A R |

16 rows × 9712 columns

```
df.describe()
```

| | year | production_value | product_emissions_MtCO2 | flaring_emissions_MtCO2 |
|---|---|---|---|---|
| **count** | 9712.000000 | 9712.000000 | 9712.000000 | 9712.000000 |
| **mean** | 1984.181322 | 360.334667 | 92.750838 | 0.504008 |
| **std** | 27.746984 | 1439.483609 | 317.321075 | 1.640646 |
| **min** | 1864.000000 | 0.007460 | 0.000399 | 0.000000 |
| **25%** | 1967.000000 | 14.383899 | 7.302732 | 0.000000 |
| **50%** | 1990.000000 | 66.557500 | 22.818983 | 0.018153 |
| **75%** | 2006.000000 | 262.262845 | 62.377203 | 0.189523 |
| **max** | 2022.000000 | 27192.000000 | 7769.222235 | 27.026872 |

```
df.isnull()
```

## 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_value |
|---|---|---|---|---|---|---|
| **3662** | 2022 | China (Coal) | Nation State | China (Coal) | Bituminous Coal | 3185.874366 |
| **3658** | 2021 | China (Coal) | Nation State | China (Coal) | Bituminous Coal | 2882.683372 |
| **3654** | 2020 | China (Coal) | Nation State | China (Coal) | Bituminous Coal | 2725.887405 |
| **3650** | 2019 | China (Coal) | Nation State | China (Coal) | Bituminous Coal | 2669.847882 |
| **3626** | 2013 | China (Coal) | Nation State | China (Coal) | Bituminous Coal | 2620.057958 |
| **...** | ... | ... | ... | ... | ... | ... |
| **8738** | 2022 | Naftogaz | State-owned Entity | Naftogaz | Oil & NGL | 0.010995 |
| **996** | 2005 | BASF | Investor-owned Company | Revus Energy | Natural Gas | 0.069529 |
| **998** | 2006 | BASF | Investor-owned Company | Revus Energy | Natural Gas | 0.066740 |
| **1000** | 2007 | BASF | Investor-owned Company | Revus Energy | Natural Gas | 0.061038 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **994** | 2004 | BASF | Investor-owned Company | Revus Energy | Natural Gas | 0.007460 |

9712 rows × 16 columns

```
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

[ ] ↳ *2 cells hidden*

## Selection by position

[ ] ↳ *3 cells hidden*