

Mhd Iqbal Pratama

105220042

1. Pertama kita harus mengecek apakah terdapat missing value, lalu setelah mengecek missing value yang ada pada kolom, kita menentukan tipe data dari kolom tersebut. Jika tipe data kolom tersebut kategori kita perlu melakukan encoder terlebih dahulu, setelah itu kita mengisi missing value dengan cara iterative imputer.

- a. Import Library dan Membaca file

```
> import pandas as pd
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer
from sklearn.preprocessing import OneHotEncoder

[118] ✓ 0.1s Python

df = pd.read_csv(r'C:\Users\LENOVO\Documents\PsdTask\volume-pengangkutan-sampah-di-kali-sungai-situ-waduk-bulan-janua

[119] ✓ 0.6s Python
```

- b. Melihat 5 data awal

```
> df.head()

[120] ✓ 0.5s Python
```

	bulan	titik lokasi	kecamatan	wilayah	panjang/luas	satuan	panjang/luas	tanggal	volume sampah perhari(m3)
0	1	Kali / Sungai Sentiong - Ruas Dewi Sri	Johar Baru	Jakarta Pusat	464		m	1	4
1	1	Kali / Sungai Sentiong - Ruas Abapon	Johar Baru	Jakarta Pusat	464		m	1	4
2	1	Kali / Sungai Sentiong - Jl. Tanah Tinggi	Johar Baru	Jakarta Pusat	464		m	1	4
3	1	Kali / Sungai Sentiong - Jl. Galur - Pombensin	Johar Baru	Jakarta Pusat	606		m	1	4
4	1	Kali / Sungai Sentiong - Jl. Timah	Johar Baru	Jakarta Pusat	310		m	1	7

- c. Mengecek missing value

```
df.info()
1] ✓ 0.1s

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 54343 entries, 0 to 54342
Data columns (total 8 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   bulan                                54343 non-null  int64
1   titik_lokasi                        54343 non-null  object
2   kecamatan                          54343 non-null  object
3   wilayah                            54343 non-null  object
4   panjang/luas                       53475 non-null  object
5   satuan_panjang/luas                32388 non-null  object
6   tanggal                            54343 non-null  int64
7   volume_sampah_perhari(m3)          54340 non-null  object
dtypes: int64(2), object(6)
memory usage: 3.3+ MB
```

- d. Mengedrop kolom yang tidak memiliki missing value

```
#drop beberapa fitur yang tidak ada missing value : titik_lokasi, kecamatan, dan bulan, wilayah dan tanggal kerna tid
df2 = df.drop(['titik_lokasi','kecamatan','bulan','wilayah','tanggal'], axis='columns')
print(df2)
[122] ✓ 0.4s Python Python

***
panjang/luas satuan_panjang/luas volume_sampah_perhari(m3)
0           464           m           4
1           464           m           4
2           464           m           4
3           606           m           4
4           310           m           7
...
54338        800           m           -
54339        700           m           -
54340       1000           m           -
54341       1000           m           -
54342        400           m           2

[54343 rows x 3 columns]
```

- e. Mengecek missing value menggunakan fungsi

```
# cek missing value dengan fungsi
def cek_null(df):
    col_na = df.isnull().sum().sort_values(ascending=False)
    percent = col_na/ len(df)*100

    missing_data = pd.concat([col_na, percent], axis=1, keys=['Total Missing Value','Percent'])
    print(missing_data[missing_data['Total Missing Value']>0])

[123] ✓ 0.4s

cek_null(df2)
[124] ✓ 0.6s

...
Total Missing Value    Percent
satuan_panjang/luas    21955  40.400788
panjang/luas           868   1.597262
volume_sampah_perhari(m3) 3   0.005520
```

f. Melakukan encoder

```
#Mengubah kategori menjadi numerik
df2['satuan_panjang/luas']=df2['satuan_panjang/luas'].astype('category')
df2['satuan_panjang/luas']=df2['satuan_panjang/luas'].cat.codes

enc = OneHotEncoder()
enc_data = pd.DataFrame(enc.fit_transform(df2[['satuan_panjang/luas']]).toarray())
df3 = df2.join(enc_data)
print(df3)

[125] ✓ 0.1s

panjang/luas  satuan_panjang/luas  volume_sampah_perhari(m3)  0  1  \
0              464                 10                4  0.0  0.0
1              464                 10                4  0.0  0.0
2              464                 10                4  0.0  0.0
3              606                 10                4  0.0  0.0
4              310                 10                7  0.0  0.0
...          ...                 ...                ...  ...  ...
54338          800                 10                -  0.0  0.0
54339          700                 10                -  0.0  0.0
54340         1000                 10                -  0.0  0.0
54341         1000                 10                -  0.0  0.0
54342          400                 10                2  0.0  0.0

2  3  4  5  6  7  8  9  10  11
```

g. Mengisi missing value menggunakan iterative imputer

```
#Mengisi missing value
imputer = IterativeImputer(random_state = 42)
imputed = imputer.fit_transform(df2)
df_imputed = pd.DataFrame(imputed, columns=df2.columns)
```

2. Data tersebut memiliki outliers karena berdasarkan grafik scatter terdapat data yang tidak terdistribusi secara tidak normal.

a. Import library dan file

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

df = pd.read_csv(r'C:\Users\LENOVO\Documents\PsdfTask\All_Diets.csv')
```

b. Melihat 5 data pertama

```
df.head()
```

	Diet_type	Recipe_name	Cuisine_type	Protein(g)	Carbs(g)	Fat(g)	Extraction_day	Extraction_time
0	paleo	Bone Broth From 'Nom Nom Paleo'	american	5.22	1.29	3.20	2022-10-16	17:20:09
1	paleo	Paleo Effect Asian-Glazed Pork Sides, A Sweet ...	south east asian	181.55	28.62	146.14	2022-10-16	17:20:09
2	paleo	Paleo Pumpkin Pie	american	30.91	302.59	96.76	2022-10-16	17:20:09
3	paleo	Strawberry Guacamole recipes	mexican	9.62	75.78	59.89	2022-10-16	17:20:09
4	paleo	Asian Cauliflower Fried "Rice" From 'Nom Nom P...	chinese	39.84	54.08	71.55	2022-10-16	17:20:09

c. Melihat deskripsi data

```
df.describe()
```

	Protein(g)	Carbs(g)	Fat(g)
count	7806.000000	7806.000000	7806.000000
mean	83.231498	152.123189	117.328542
std	89.797282	185.907322	122.098117
min	0.000000	0.060000	0.000000
25%	24.415000	36.162500	41.067500
50%	56.280000	93.415000	84.865000
75%	112.357500	205.915000	158.290000
max	1273.610000	3405.550000	1930.240000

d. Mengecek Null data

```
#Cek data Null

print('Null Data', 'Total')
print(df.isnull().sum())
```

✓ 0.6s

Null Data	Total
Diet_type	0
Recipe_name	0
Cuisine_type	0
Protein(g)	0
Carbs(g)	0
Fat(g)	0
Extraction_day	0
Extraction_time	0

dtype: int64

e. Scatting

