ML EXP-1

DATE:

Name: Prerna Jadhav
SapId: 60004220127
Batch: C22
Branch: Computer Engineering
course: Machine learning

AIM: To perform data preprocessing in terms of handling, missing data, removing outliers, eliminating duplicate rows & modifying the datatypes.

DESCRIPTION OF EXPERIMENT: Dirty data simply means data that is erroroeus. Duplicacy of records incomplete or outdated data and improper parsing can make data dirty. This data needs to be cleared. Data cleaning refers to the process of 'cleaning' this data, by identifying errors in the data and then rectifying them. It is an important step, is an basic necessity in the Machine learning projects.

Cleaning data in Python: Missing value - calculate the percentage of values missing in each column and then storing this information in Dataframe. Drop observation: one way could be to drop those observation that contains any null value in them for any of the columns. This works when the percentage of missing values in each column is very less.

Removing columns (features): Another way to tackle missing values in a dataset would be to drop those columns as features that have a significant percentange of missing value.

Impute missing values: Impute the missing value in each numerical column with median value of the column.

Outliers: It is an usual observation that lies away from the majority of the data. Outliers can affect the performance of a ML model significantly.

Duplicate records: Data can sometimes contain duplicate values & it is important to remove them. since in our dataset we have a few duplicates we will drop them

fixing data type: Often in the dataset, values are not stored in the correct data type. This can create a problem in later stages and may not get the desired o/p or may get errors.

CONCLUSION: Thus, we cleaned the dataset & learnt its importance in the data processing.

- Name: Prerna Sunil Jadhav
- Sap ID: 60004220127
- Batch: C22
- Branch: Computer Engineering
- Course: Machine Learning
- Experiment 1: Data Cleaning

```
import pandas as pd
import numpy as np

import warnings

# Filter or ignore specific warning messages
warnings.filterwarnings("ignore")

# loading data frame
country_data=pd.read_csv("../content/countries_of_the_world.csv")
print(country_data.dtypes,"\n")
print(country_data.columns,"\n")
country_data.head()
```

```
Country                               object
Region                                object
Population                             int64
Area (sq. mi.)                         int64
Pop. Density (per sq. mi.)            object
Coastline (coast/area ratio)          object
Net migration                         object
Infant mortality (per 1000 births)    object
GDP ($ per capita)                   float64
Literacy (%)                          object
Phones (per 1000)                     object
Arable (%)                            object
Crops (%)                             object
Other (%)                             object
Climate                               object
Birthrate                             object
Deathrate                             object
Agriculture                           object
Industry                              object
Service                               object
dtype: object

Index(['Country', 'Region', 'Population', 'Area (sq. mi.)',
       'Pop. Density (per sq. mi.)', 'Coastline (coast/area ratio)',
       'Net migration', 'Infant mortality (per 1000 births)',
       'GDP ($ per capita)', 'Literacy (%)', 'Phones (per 1000)', 'Arable (%)',
       'Crops (%)', 'Other (%)', 'Climate', 'Birthrate', 'Deathrate',
       'Agriculture', 'Industry', 'Service'],
      dtype='object')
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	In mort (per bi
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	48,0	0,00	23,06	1
1	Albania	EASTERN EUROPE	3581655	28748	124,6	1,26	-4,93	
2	Algeria	NORTHERN AFRICA	32930091	2381740	13,8	0,04	-0,39	
3	American Samoa	OCEANIA	57794	199	290,4	58,29	-20,71	
4	Andorra	WESTERN EUROPE	71201	468	152,1	0,00	6,6	

```
print(country_data.info())
print(country_data.describe())
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 179 entries, 0 to 226
Data columns (total 20 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Country                             179 non-null    object
 1   Region                              179 non-null    object
 2   Population                          179 non-null    int64
 3   Area (sq. mi.)                      179 non-null    int64
 4   Pop. Density (per sq. mi.)          179 non-null    float64
 5   Coastline (coast/area ratio)        179 non-null    float64
 6   Net migration                       179 non-null    float64
 7   Infant mortality (per 1000 births)  179 non-null    float64
 8   GDP ($ per capita)                  179 non-null    float64
 9   Literacy (%)                        179 non-null    float64
 10  Phones (per 1000)                   179 non-null    float64
 11  Arable (%)                          179 non-null    float64
 12  Crops (%)                           179 non-null    float64
 13  Other (%)                           179 non-null    float64
 14  Climate                             179 non-null    object
 15  Birthrate                           179 non-null    float64
 16  Deathrate                           179 non-null    float64
 17  Agriculture                         179 non-null    float64
 18  Industry                            179 non-null    float64
 19  Service                             179 non-null    float64
dtypes: float64(15), int64(2), object(3)
memory usage: 29.4+ KB
None
         Population  Area (sq. mi.)  Pop. Density (per sq. mi.)  \
count  1.790000e+02    1.790000e+02                  179.000000
mean   3.421415e+07    5.641830e+05                 2948.050279
std    1.317639e+08    1.395657e+06                13793.525570
min    1.347700e+04    2.800000e+01                   18.000000
25%    1.188580e+06    1.991500e+04                  268.000000
50%    6.940432e+06    1.184800e+05                  669.000000
75%    2.086014e+07    4.964410e+05                 1647.000000
max    1.313974e+09    9.631420e+06               161830.000000

       Coastline (coast/area ratio)  Net migration  \
count                    179.000000     179.000000
mean                    1649.519553      -5.972067
std                     7397.760059     463.095246
min                        0.000000   -2099.000000
25%                        9.000000     -79.000000
50%                       63.000000       0.000000
75%                      535.500000      27.000000
max                    87066.000000    2306.000000

       Infant mortality (per 1000 births)  GDP ($ per capita)  Literacy (%)  \
count                          179.000000          179.000000    179.000000
mean                          3461.083799         9125.698324    819.441341
std                           3513.033899         9644.123141    198.375370
min                             31.000000          500.000000    176.000000
25%                            741.500000         1800.000000    699.500000
50%                           2055.000000         5100.000000    909.000000
75%                           5533.500000        12950.000000    978.000000
max                          16307.000000        37800.000000   1000.000000
```

```
duplicate_rows = country_data.duplicated()
print("Number of duplicate rows:", duplicate_rows.sum())
```

```
Number of duplicate rows: 0
```

```
# Changing data type from str to float
column_to_float=['Pop. Density (per sq. mi.)','Coastline (coast/area ratio)',
                 'Net migration', 'Infant mortality (per 1000 births)',
                 'Literacy (%)', 'Phones (per 1000)', 'Arable (%)', 'Crops (%)',
                 'Other (%)', 'Birthrate', 'Deathrate', 'Agriculture',
                 'Industry', 'Service']

for column in column_to_float:
    country_data[column]=country_data[column].astype(str)
    country_data[column]=country_data[column].str.replace(",","")
    country_data[column]=country_data[column].str.replace("$","")
    country_data[column]=country_data[column].str.replace("%","").astype(float)

print(country_data.dtypes,"\n")
country_data.head()
```

```
Country                               object
Region                                object
Population                             int64
Area (sq. mi.)                         int64
Pop. Density (per sq. mi.)           float64
Coastline (coast/area ratio)         float64
Net migration                        float64
Infant mortality (per 1000 births)   float64
GDP ($ per capita)                   float64
Literacy (%)                         float64
Phones (per 1000)                    float64
Arable (%)                           float64
Crops (%)                            float64
Other (%)                            float64
Climate                               object
Birthrate                            float64
Deathrate                            float64
Agriculture                          float64
Industry                             float64
Service                              float64
dtype: object
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	In mort (per bi
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	480.0	0.0	2306.0	16
1	Albania	EASTERN EUROPE	3581655	28748	1246.0	126.0	-493.0	2
2	Algeria	NORTHERN AFRICA	32930091	2381740	138.0	4.0	-39.0	
3	American Samoa	OCEANIA	57794	199	2904.0	5829.0	-2071.0	
4	Andorra	WESTERN EUROPE	71201	468	1521.0	0.0	66.0	

```
country_data = country_data.dropna()
print(country_data)
```

```
                Country                               Region  Population  \
0           Afghanistan         ASIA (EX. NEAR EAST)             31056997
1               Albania   EASTERN EUROPE                           3581655
2               Algeria   NORTHERN AFRICA                         32930091
6              Anguilla               LATIN AMER. & CARIB           13477
7     Antigua & Barbuda               LATIN AMER. & CARIB           69108
..                   ...                                  ...         ...
218           Venezuela               LATIN AMER. & CARIB        25730435
219             Vietnam         ASIA (EX. NEAR EAST)             84402966
224               Yemen   NEAR EAST                               21456188
225              Zambia   SUB-SAHARAN AFRICA                      11502010
226            Zimbabwe   SUB-SAHARAN AFRICA                      12236805

     Area (sq. mi.)  Pop. Density (per sq. mi.)  Coastline (coast/area ratio)  \
0            647500                       480.0                           0.0
```

```
1            28748                   1246.0                        126.0
2          2381740                    138.0                          4.0
6              102                   1321.0                       5980.0
7              443                   1560.0                       3454.0
..             ...                      ...                          ...
218         912050                    282.0                         31.0
219         329560                   2561.0                        105.0
224         527970                    406.0                         36.0
225         752614                    153.0                          0.0
226         390580                    313.0                          0.0

     Net migration  Infant mortality (per 1000 births)  GDP ($ per capita)  \
0           2306.0                            16307.0               700.0
1           -493.0                             2152.0              4500.0
2            -39.0                               31.0              6000.0
6           1076.0                             2103.0              8600.0
7           -615.0                             1946.0             11000.0
..             ...                                ...                 ...
218           -4.0                              222.0              4800.0
219          -45.0                             2595.0              2500.0
224            0.0                              615.0               800.0
225            0.0                             8829.0               800.0
226            0.0                             6769.0              1900.0

     Literacy (%)  Phones (per 1000)  Arable (%)  Crops (%)  Other (%)  \
0           360.0               32.0      1213.0       22.0     8765.0
1           865.0              712.0      2109.0      442.0     7449.0
2           700.0              781.0       322.0       25.0     9653.0
6           950.0             4600.0         0.0        0.0      100.0
7           890.0             5499.0      1818.0      455.0     7727.0
..            ...                ...         ...        ...        ...
218         934.0             1401.0       295.0       92.0     9613.0
219         903.0             1877.0      1997.0      595.0     7408.0
224         502.0              372.0       278.0       24.0     9698.0
225         806.0               82.0       708.0        3.0      929.0
226         907.0              268.0       832.0       34.0     9134.0

    Climate  Birthrate  Deathrate  Agriculture  Industry  Service
0         1      466.0     2034.0         38.0      24.0     38.0
1         3     1511.0      522.0        232.0     188.0    579.0
2         1     1714.0      461.0        101.0       6.0    298.0
6         2     1417.0      534.0          4.0      18.0     78.0
7         2     1693.0      537.0         38.0      22.0    743.0
```

```
country_data.isna()
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	c
0	False	False	False	False	False	False	False	False	
1	False	False	False	False	False	False	False	False	
2	False	False	False	False	False	False	False	False	
6	False	False	False	False	False	False	False	False	
7	False	False	False	False	False	False	False	False	
...	...	...	...	...	...	...	...	...	
218	False	False	False	False	False	False	False	False	
219	False	False	False	False	False	False	False	False	
224	False	False	False	False	False	False	False	False	
225	False	False	False	False	False	False	False	False	
226	False	False	False	False	False	False	False	False	

179 rows × 20 columns

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()
cols_to_normalize = ['Net migration']
scaled_data = scaler.fit(country_data[cols_to_normalize])
country_data[cols_to_normalize] = scaler.transform(country_data[cols_to_normalize])
country_data.head()
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	In mort (per bi
0	Afghanistan	ASIA (EX. NEAR EAST)	31056997	647500	480.0	0.0	5.006437	16
1	Albania	EASTERN EUROPE	3581655	28748	1246.0	126.0	-1.054630	2
2	Algeria	NORTHERN AFRICA	32930091	2381740	138.0	4.0	-0.071520	
6	Anguilla	LATIN AMER. & CARIB	13477	102	1321.0	5980.0	2.342946	2
7	Antigua & Barbuda	LATIN AMER. & CARIB	69108	443	1560.0	3454.0	-1.318814	1

```
country_data.to_csv('cleaned_data.csv', index=False)
```