Exno:1

Data Cleaning Process

AIM

To read the given data and perform data cleaning and save the cleaned data to a file.

Explanation

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect ,incompleted , irrelevant , duplicated or improperly formatted. Data cleaning is not simply about erasing data ,but rather finding a way to maximize datasets accuracy without necessarily deleting the information.

Algorithm

STEP 1: Read the given Data
STEP 2: Get the information about the data
STEP 3: Remove the null values from the data
STEP 4: Save the Clean data to the file
STEP 5: Remove outliers using IQR
STEP 6: Use zscore of to remove outliers

Coding and Output:

Data cleaning process:

```
import pandas as pd
df=pd.read_csv("/content/SAMPLEIDS.csv")
df
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
0	1	1220121	ARUN	2000-02-10	MALE	THANDALAM	82.0	81.0	90.0	NaN	NaN	NaN
1	2	1220122	BABU	1999-01-25	MALE	KANCHIPURAM	56.0	61.0	80.0	56.0	253.0	84.333333
2	3	1220123	CHARAN	2000.09.21	MALE	THANDALAM	NaN	59.0	60.0	70.0	NaN	0.000000
3	4	1220124	DEVA	2000-11-09	MALE	POONAMALEE	74.0	79.0	80.0	74.0	307.0	102.333333
4	5	1220125	ESTER	2000-11-21	FEMALE	CHITHUR	92.0	95.0	96.0	92.0	375.0	125.000000
5	6	1220126	FARHANA	1999-03-05	FEMALE	THANDALAM	91.0	88.0	90.0	91.0	360.0	120.000000
6	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
7	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
8	8	1220128	HEMA	1999-01-25	FEMALE	POONAMALEE	95.0	96.0	90.0	95.0	376.0	125.333333
9	9	1220129	INDRA	2000.09.21	FEMALE	KANCHIPURAM	64.0	NaN	NaN	64.0	NaN	0.000000
10	10	1220130	JAHITH	2000-11-09	MALE	THANDALAM	34.0	45.0	50.0	34.0	163.0	54.333333
11	11	1220131	KANI	2000-11-21	FEMALE	CHITHUR	96.0	95.0	96.0	96.0	383.0	127.666667
12	12	1220132	LATHESSH	1999-03-05	MALE	THANDALAM	NaN	68.0	70.0	70.0	208.0	69.333333
13	13	1220133	MANI	2000-10-02	MALE	KANCHIPURAM	71.0	76.0	NaN	71.0	NaN	0.000000
14	14	1220134	NANI	20001109	MALE	POONAMALEE	79.0	77.0	80.0	79.0	315.0	105.000000
15	15	1220135	NaN	19990125	NaN	NaN	NaN	NaN	NaN	NaN	0.0	0.000000
16	16	1220136	PRATHAP	20000921	MALE	KANCHIPURAM	86.0	84.0	90.0	86.0	346.0	115.333333
17	17	1220137	RAGHU	20001109	MALE	POONAMALEE	67.0	64.0	70.0	NaN	201.0	67.000000
18	18	1220138	RATHI	20001121	FEMALE	KANCHIPURAM	81.0	86.0	90.0	81.0	338.0	112.666667
19	19	1220139	SARVESH	19990305	MALE	THANDALAM	84.0	87.0	NaN	84.0	NaN	0.000000
20	20	1220140	SANTHOSH	20001002	MALE	KANCHIPURAM	76.0	69.0	80.0	76.0	301.0	100.333333

```
df.head()
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
0	1	1220121	ARUN	2000-02-10	MALE	THANDALAM	82.0	81.0	90.0	NaN	NaN	NaN
1	2	1220122	BABU	1999-01-25	MALE	KANCHIPURAM	56.0	61.0	80.0	56.0	253.0	84.333333
2	3	1220123	CHARAN	2000.09.21	MALE	THANDALAM	NaN	59.0	60.0	70.0	NaN	0.000000
3	4	1220124	DEVA	2000-11-09	MALE	POONAMALEE	74.0	79.0	80.0	74.0	307.0	102.333333
4	5	1220125	ESTER	2000-11-21	FEMALE	CHITHUR	92.0	95.0	96.0	92.0	375.0	125.000000

```
df.tail(5)
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
16	16	1220136	PRATHAP	20000921	MALE	KANCHIPURAM	86.0	84.0	90.0	86.0	346.0	115.333333
17	17	1220137	RAGHU	20001109	MALE	POONAMALEE	67.0	64.0	70.0	NaN	201.0	67.000000
18	18	1220138	RATHI	20001121	FEMALE	KANCHIPURAM	81.0	86.0	90.0	81.0	338.0	112.666667
19	19	1220139	SARVESH	19990305	MALE	THANDALAM	84.0	87.0	NaN	84.0	NaN	0.000000
20	20	1220140	SANTHOSH	20001002	MALE	KANCHIPURAM	76.0	69.0	80.0	76.0	301.0	100.333333

```
df.isnull()
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
0	False	False	False	False	False	False	False	False	False	True	True	True
1	False	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	True	False	False	False	True	False
3	False	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False	False
5	False	False	False	False	False	False	False	False	False	False	False	False
6	False	False	False	False	False	False	False	False	False	False	False	False
7	False	False	False	False	False	False	False	False	False	False	False	False
8	False	False	False	False	False	False	False	False	False	False	False	False
9	False	False	False	False	False	False	False	True	True	False	True	False
10	False	False	False	False	False	False	False	False	False	False	False	False
11	False	False	False	False	False	False	False	False	False	False	False	False
12	False	False	False	False	False	False	True	False	False	False	False	False
13	False	False	False	False	False	False	False	False	True	False	True	False
14	False	False	False	False	False	False	False	False	False	False	False	False
15	False	False	True	False	True	True	True	True	True	True	False	False
16	False	False	False	False	False	False	False	False	False	False	False	False
17	False	False	False	False	False	False	False	False	False	True	False	False
18	False	False	False	False	False	False	False	False	False	False	False	False
19	False	False	False	False	False	False	False	False	True	False	True	False
20	False	False	False	False	False	False	False	False	False	False	False	False

```
df.notnull()
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
0	True	True	True	True	True	True	True	True	True	False	False	False
1	True	True	True	True	True	True	True	True	True	True	True	True
2	True	True	True	True	True	True	False	True	True	True	False	True
3	True	True	True	True	True	True	True	True	True	True	True	True
4	True	True	True	True	True	True	True	True	True	True	True	True
5	True	True	True	True	True	True	True	True	True	True	True	True
6	True	True	True	True	True	True	True	True	True	True	True	True
7	True	True	True	True	True	True	True	True	True	True	True	True
8	True	True	True	True	True	True	True	True	True	True	True	True
9	True	True	True	True	True	True	True	False	False	True	False	True
10	True	True	True	True	True	True	True	True	True	True	True	True
11	True	True	True	True	True	True	True	True	True	True	True	True
12	True	True	True	True	True	True	False	True	True	True	True	True
13	True	True	True	True	True	True	True	True	False	True	False	True
14	True	True	True	True	True	True	True	True	True	True	True	True
15	True	True	False	True	False	False	False	False	False	False	True	True
16	True	True	True	True	True	True	True	True	True	True	True	True
17	True	True	True	True	True	True	True	True	True	False	True	True
18	True	True	True	True	True	True	True	True	True	True	True	True
19	True	True	True	True	True	True	True	True	False	True	False	True
20	True	True	True	True	True	True	True	True	True	True	True	True

```
df.dropna(axis=0)
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
1	2	1220122	BABU	1999-01-25	MALE	KANCHIPURAM	56.0	61.0	80.0	56.0	253.0	84.333333
3	4	1220124	DEVA	2000-11-09	MALE	POONAMALEE	74.0	79.0	80.0	74.0	307.0	102.333333
4	5	1220125	ESTER	2000-11-21	FEMALE	CHITHUR	92.0	95.0	96.0	92.0	375.0	125.000000
5	6	1220126	FARHANA	1999-03-05	FEMALE	THANDALAM	91.0	88.0	90.0	91.0	360.0	120.000000
6	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
7	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
8	8	1220128	HEMA	1999-01-25	FEMALE	POONAMALEE	95.0	96.0	90.0	95.0	376.0	125.333333
10	10	1220130	JAHITH	2000-11-09	MALE	THANDALAM	34.0	45.0	50.0	34.0	163.0	54.333333
11	11	1220131	KANI	2000-11-21	FEMALE	CHITHUR	96.0	95.0	96.0	96.0	383.0	127.666667
14	14	1220134	NANI	20001109	MALE	POONAMALEE	79.0	77.0	80.0	79.0	315.0	105.000000
16	16	1220136	PRATHAP	20000921	MALE	KANCHIPURAM	86.0	84.0	90.0	86.0	346.0	115.333333
18	18	1220138	RATHI	20001121	FEMALE	KANCHIPURAM	81.0	86.0	90.0	81.0	338.0	112.666667
20	20	1220140	SANTHOSH	20001002	MALE	KANCHIPURAM	76.0	69.0	80.0	76.0	301.0	100.333333

```
df.dropna(axis=1)
```

	SNO	REGNO	DOB
0	1	1220121	2000-02-10
1	2	1220122	1999-01-25
2	3	1220123	2000.09.21
3	4	1220124	2000-11-09
4	5	1220125	2000-11-21
5	6	1220126	1999-03-05
6	7	1220127	2000-10-02
7	7	1220127	2000-10-02
8	8	1220128	1999-01-25
9	9	1220129	2000.09.21
10	10	1220130	2000-11-09
11	11	1220131	2000-11-21
12	12	1220132	1999-03-05
13	13	1220133	2000-10-02
14	14	1220134	20001109
15	15	1220135	19990125
16	16	1220136	20000921
17	17	1220137	20001109
18	18	1220138	20001121
19	19	1220139	19990305
20	20	1220140	20001002

```
df.fillna(0)
```

	SNO	REGNO	NAME	DOB	GENDER	ADDRESS	M1	M2	M3	M4	TOTAL	AVG
0	1	1220121	ARUN	2000-02-10	MALE	THANDALAM	82.0	81.0	90.0	0.0	0.0	0.000000
1	2	1220122	BABU	1999-01-25	MALE	KANCHIPURAM	56.0	61.0	80.0	56.0	253.0	84.333333
2	3	1220123	CHARAN	2000.09.21	MALE	THANDALAM	0.0	59.0	60.0	70.0	0.0	0.000000
3	4	1220124	DEVA	2000-11-09	MALE	POONAMALEE	74.0	79.0	80.0	74.0	307.0	102.333333
4	5	1220125	ESTER	2000-11-21	FEMALE	CHITHUR	92.0	95.0	96.0	92.0	375.0	125.000000
5	6	1220126	FARHANA	1999-03-05	FEMALE	THANDALAM	91.0	88.0	90.0	91.0	360.0	120.000000
6	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
7	7	1220127	GANI	2000-10-02	MALE	KANCHIPURAM	49.0	51.0	70.0	49.0	219.0	73.000000
8	8	1220128	HEMA	1999-01-25	FEMALE	POONAMALEE	95.0	96.0	90.0	95.0	376.0	125.333333
9	9	1220129	INDRA	2000.09.21	FEMALE	KANCHIPURAM	64.0	0.0	0.0	64.0	0.0	0.000000
10	10	1220130	JAHITH	2000-11-09	MALE	THANDALAM	34.0	45.0	50.0	34.0	163.0	54.333333
11	11	1220131	KANI	2000-11-21	FEMALE	CHITHUR	96.0	95.0	96.0	96.0	383.0	127.666667
12	12	1220132	LATHESSH	1999-03-05	MALE	THANDALAM	0.0	68.0	70.0	70.0	208.0	69.333333
13	13	1220133	MANI	2000-10-02	MALE	KANCHIPURAM	71.0	76.0	0.0	71.0	0.0	0.000000
14	14	1220134	NANI	20001109	MALE	POONAMALEE	79.0	77.0	80.0	79.0	315.0	105.000000
15	15	1220135	0	19990125	0	0	0.0	0.0	0.0	0.0	0.0	0.000000
16	16	1220136	PRATHAP	20000921	MALE	KANCHIPURAM	86.0	84.0	90.0	86.0	346.0	115.333333
17	17	1220137	RAGHU	20001109	MALE	POONAMALEE	67.0	64.0	70.0	0.0	201.0	67.000000
18	18	1220138	RATHI	20001121	FEMALE	KANCHIPURAM	81.0	86.0	90.0	81.0	338.0	112.666667
19	19	1220139	SARVESH	19990305	MALE	THANDALAM	84.0	87.0	0.0	84.0	0.0	0.000000
20	20	1220140	SANTHOSH	20001002	MALE	KANCHIPURAM	76.0	69.0	80.0	76.0	301.0	100.333333

```
print(df.shape)
```

```
(21, 12)
```

IQR:

```
import pandas as pd
import seaborn as sns
ir=pd.read_csv('/content/iris.csv')
ir
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

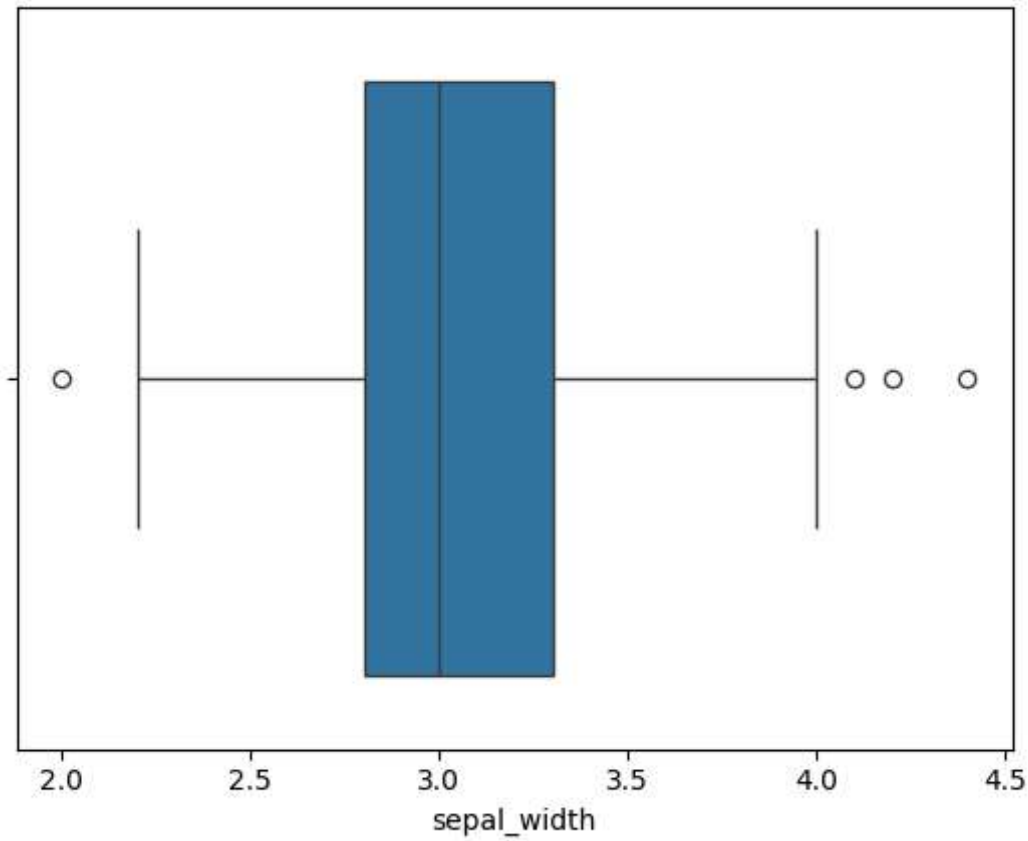
150 rows × 5 columns

```
ir.describe()
```

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

```
sns.boxplot(x='sepal_width',data=ir)
```

```
<Axes: xlabel='sepal_width'>
```




```
c1=ir.sepal_width.quantile(0.25)
c3=ir.sepal_width.quantile(0.75)
iq=c3-c1
print(c3)
```

```
3.3
```

```
rid=ir[((ir.sepal_width<(c1-1.5*iq))|(ir.sepal_width>(c3+1.5*iq)))]
rid['sepal_width']
```

	sepal_width
15	4.4
32	4.1
33	4.2
60	2.0

dtype: float64

```
delid=ir[~((ir.sepal_width<(c1-1.5*iq))|(ir.sepal_width>(c3+1.5*iq)))]
delid
```

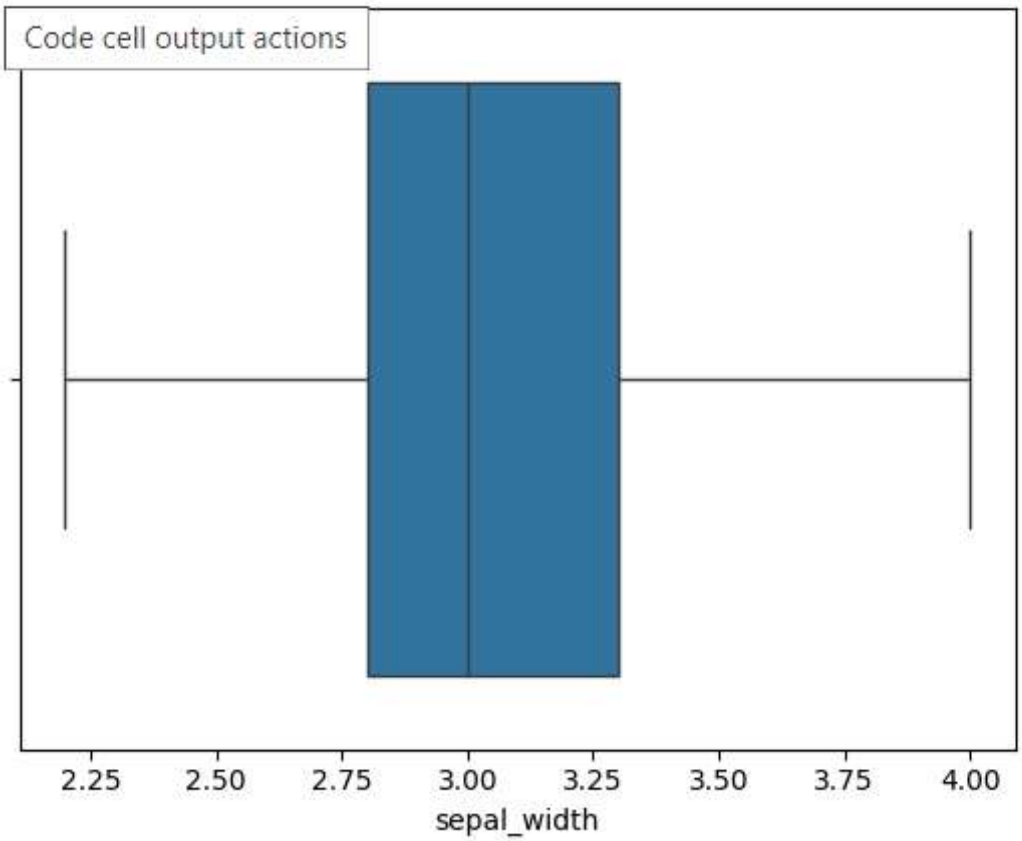
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

146 rows × 5 columns

```
sns.boxplot(x='sepal_width',data=delid)
```

```
<Axes: xlabel='sepal_width'>
```

Code cell output actions



```
import matplotlib.pyplot as plt
import pandas as pd
import pandas as pd
import numpy as np
import scipy.stats as stats
dataset=pd.read_csv("/content/heights.csv")
dataset
```

	name	height
0	mohan	5.9
1	maria	5.2
2	sakib	5.1
3	tao	5.5
4	virat	4.9
5	khusbu	5.4
6	dmitry	6.2
7	selena	6.5
8	john	7.1
9	imran	14.5
10	jose	6.1
11	deepika	5.6
12	yoseph	1.2
13	binod	5.5

```
df = pd.read_csv("heights.csv")
q1 = df['height'].quantile(0.25)
q2 = df['height'].quantile(0.5)
q3 = df['height'].quantile(0.75)
iqr = q3-q1
iqr
```

```
0.9249999999999998
```

```
low = q1 - 1.5*iqr
low
```

```
3.8625000000000003
```

```
high = q3 + 1.5*iqr
high
```

```
7.5625
```

```
df1 = df[((df['height'] >=low)& (df['height'] <=high))]
df1
```

	name	height
0	mohan	5.9
1	maria	5.2
2	sakib	5.1
3	tao	5.5
4	virat	4.9
5	khusbu	5.4
6	dmitry	6.2
7	selena	6.5
8	john	7.1
10	jose	6.1
11	deepika	5.6
13	binod	5.5

```
z = np.abs(stats.zscore(df['height']))
z
```

	height
0	0.055998
1	0.317320
2	0.354652
3	0.205325
4	0.429315
5	0.242656
6	0.055998
7	0.167993
8	0.391983
9	3.154532
10	0.018666
11	0.167993
12	1.810589
13	0.205325

dtype: float64

```
df1 = df[z<3]
df1
```

	name	height
0	mohan	5.9
1	maria	5.2
2	sakib	5.1
3	tao	5.5
4	virat	4.9
5	khusbu	5.4
6	dmitry	6.2
7	selena	6.5
8	john	7.1
10	jose	6.1
11	deepika	5.6
12	yoseph	1.2
13	binod	5.5

Result

Thus the Data Cleaning Process and Detecting and Removal of Outliers is executed successfully.