Data Cleaning Day1

July 27, 2023

[1]:
```
import numpy as np
import pandas as pd
```

[71]:
```
df=pd.read_csv("C:\\Users\\ASUS\\Desktop\\DS Course\\Data Cleaning\\iris - Missing.csv")
```

[6]:
```
df.head(15)
```

[6]:

	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	NaN	NaN	NaN	NaN	NaN
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	NaN	setosa
8	4.4	2.9	1.4	NaN	NaN
9	4.9	3.1	1.5	NaN	setosa
10	NaN	3.7	1.5	NaN	setosa
11	4.8	3.4	1.6	NaN	setosa
12	4.8	3.0	1.4	NaN	setosa
13	4.3	3.0	1.1	NaN	setosa
14	NaN	NaN	NaN	NaN	setosa

[ ]:
```
#removing Columns
```

[12]:
```
df.isna()                    #show are there any null values in the cell
```

[12]:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	False	False	False	False	False
1	False	False	False	False	False
2	True	True	True	True	True
3	False	False	False	False	False
4	False	False	False	False	False
..	…	…	…	…	…
145	False	False	False	True	False
146	False	False	False	True	False

```
147         False        False         False         True    False
148         False        False         False         True    False
149         False        False         False        False    False

[150 rows x 5 columns]
```

[10]:
```
df.isna().sum()            #shows how many null values are there in a column
```

[10]:
```
Sepal.Length     5
Sepal.Width      7
Petal.Length     2
Petal.Width     85
Species          5
dtype: int64
```

[13]:
```
df.isna().sum(axis=1).values
```

[13]:
```
array([0, 0, 5, 0, 0, 0, 0, 1, 2, 1, 2, 1, 1, 1, 4, 1, 0, 0, 0, 1, 0, 0,
       0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 1, 0, 0, 0, 1, 0, 0, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1,
       1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
       1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0], dtype=int64)
```

[14]:
```
df.isna().any()              #shows are the any null values in the column
```

[14]:
```
Sepal.Length    True
Sepal.Width     True
Petal.Length    True
Petal.Width     True
Species         True
dtype: bool
```

[15]:
```
df.isna().any(axis=1).values
```

[15]:
```
array([False, False,  True, False, False, False, False,  True,  True,
        True,  True,  True,  True,  True,  True,  True, False, False,
       False,  True, False, False, False, False,  True, False, False,
       False, False, False, False, False, False, False,  True,
       False, False, False, False, False, False, False, False, False,
        True, False, False, False,  True, False, False,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True, False, False, False,  True, False,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True,  True,  True,  True,  True,
```

```
        True,  True,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True, False, False, False, False, False, False,
       False, False, False, False, False, False, False, False, False,
       False, False,  True,  True,  True,  True,  True,  True,  True,
        True,  True,  True,  True,  True, False])
```

[16]:
```
df.isna().all()       #return True if all the values in df are null or missing,␣
 ↪and False otherwise.
```

[16]:
```
Sepal.Length    False
Sepal.Width     False
Petal.Length    False
Petal.Width     False
Species         False
dtype: bool
```

[17]:
```
df.isna().all(axis=1)       #return a boolean Series where each value is True
                            #if all the values in the corresponding row of df␣
 ↪are null or missing,
                            #and False otherwise.
```

[17]:
```
0      False
1      False
2       True
3      False
4      False
       …
145    False
146    False
147    False
148    False
149    False
Length: 150, dtype: bool
```

[18]:
```
df.isna().sum()/len(df)*100         #give the missing data percentage of each␣
 ↪column
```

[18]:
```
Sepal.Length     3.333333
Sepal.Width      4.666667
Petal.Length     1.333333
Petal.Width     56.666667
Species          3.333333
dtype: float64
```

[72]:
```
df.drop("Petal.Width",axis=1,inplace=True) #delete the Petal.Width column␣
 ↪because it has more missing data
df
```

[72]:
```
     Sepal.Length  Sepal.Width  Petal.Length    Species
0             5.1          3.5           1.4     setosa
1             4.9          3.0           1.4     setosa
2             NaN          NaN           NaN        NaN
3             4.6          3.1           1.5     setosa
4             5.0          3.6           1.4     setosa
..            …            …             …       …
145           6.7          3.0           5.2  virginica
146           6.3          2.5           5.0  virginica
147           6.5          3.0           5.2  virginica
148           6.2          3.4           5.4  virginica
149           5.9          3.0           5.1  virginica

[150 rows x 4 columns]
```

[22]:
```
df.head(15)
```

[22]:
```
    Sepal.Length  Sepal.Width  Petal.Length Species
0            5.1          3.5           1.4  setosa
1            4.9          3.0           1.4  setosa
2            NaN          NaN           NaN     NaN
3            4.6          3.1           1.5  setosa
4            5.0          3.6           1.4  setosa
5            5.4          3.9           1.7  setosa
6            4.6          3.4           1.4  setosa
7            5.0          3.4           1.5  setosa
8            4.4          2.9           1.4     NaN
9            4.9          3.1           1.5  setosa
10           NaN          3.7           1.5  setosa
11           4.8          3.4           1.6  setosa
12           4.8          3.0           1.4  setosa
13           4.3          3.0           1.1  setosa
14           NaN          NaN           NaN  setosa
```

[23]:
```
df1=df.dropna(how="any")                # remove a row if it has at least one␣
 ↪missing data (row 2,10,14 removed)
df1.head(15)
```

[23]:
```
    Sepal.Length  Sepal.Width  Petal.Length Species
0            5.1          3.5           1.4  setosa
1            4.9          3.0           1.4  setosa
3            4.6          3.1           1.5  setosa
4            5.0          3.6           1.4  setosa
5            5.4          3.9           1.7  setosa
6            4.6          3.4           1.4  setosa
7            5.0          3.4           1.5  setosa
9            4.9          3.1           1.5  setosa
```

```
11           4.8          3.4           1.6  setosa
12           4.8          3.0           1.4  setosa
13           4.3          3.0           1.1  setosa
15           5.7          4.4           1.5  setosa
16           5.4          3.9           1.3  setosa
17           5.1          3.5           1.4  setosa
18           5.7          3.8           1.7  setosa
```

[24]:
```
df1=df.dropna(how="all")        #only removes rows if all the columns are␣
 ↪missing data of that row.(row 2 removed)
df1.head(15)
```

[24]:
```
    Sepal.Length  Sepal.Width  Petal.Length Species
0            5.1          3.5           1.4  setosa
1            4.9          3.0           1.4  setosa
3            4.6          3.1           1.5  setosa
4            5.0          3.6           1.4  setosa
5            5.4          3.9           1.7  setosa
6            4.6          3.4           1.4  setosa
7            5.0          3.4           1.5  setosa
8            4.4          2.9           1.4     NaN
9            4.9          3.1           1.5  setosa
10           NaN          3.7           1.5  setosa
11           4.8          3.4           1.6  setosa
12           4.8          3.0           1.4  setosa
13           4.3          3.0           1.1  setosa
14           NaN          NaN           NaN  setosa
15           5.7          4.4           1.5  setosa
```

[ ]:
```
# removing Rows
```

[33]:
```
df.notnull().sum()          #get the all the not null data
```

[33]:
```
Sepal.Length    145
Sepal.Width     143
Petal.Length    148
Species         145
dtype: int64
```

[34]:
```
df.notna().sum()
```

[34]:
```
Sepal.Length    145
Sepal.Width     143
Petal.Length    148
Species         145
dtype: int64
```

[35]:
```
df.notna().sum(axis=1).values       #gives how many not null values in each row.
```

[35]:
```
array([4, 4, 0, 4, 4, 4, 4, 4, 3, 4, 3, 4, 4, 4, 1, 4, 4, 4, 4, 3, 4, 4,
       4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 3, 4, 4, 4, 3, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4,
       4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4], dtype=int64)
```

[42]:
```
df2=df.dropna(thresh=3)  #drops all rows from a DataFrame df that have less than three non-null values.
df2.head(15)
```

[42]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	NaN
9	4.9	3.1	1.5	setosa
10	NaN	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
15	5.7	4.4	1.5	setosa
16	5.4	3.9	1.3	setosa

[44]:
```
df2.notna().sum(axis=1).values
```

[44]:
```
array([4, 4, 4, 4, 4, 4, 4, 3, 4, 3, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4,
       3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3,
       4, 4, 4, 3, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 4, 4, 4,
       4, 4, 4, 4, 4, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4], dtype=int64)
```

[40]:
```
df3=df.dropna(thresh=4) #drops all rows from a DataFrame df that have less than four non-null values.
df3.head(15)
```

[40]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
3	4.6	3.1	1.5	setosa

```
4             5.0           3.6           1.4  setosa
5             5.4           3.9           1.7  setosa
6             4.6           3.4           1.4  setosa
7             5.0           3.4           1.5  setosa
9             4.9           3.1           1.5  setosa
11            4.8           3.4           1.6  setosa
12            4.8           3.0           1.4  setosa
13            4.3           3.0           1.1  setosa
15            5.7           4.4           1.5  setosa
16            5.4           3.9           1.3  setosa
17            5.1           3.5           1.4  setosa
18            5.7           3.8           1.7  setosa
```

[45]:
```
df3.notna().sum(axis=1).values
```

[45]:
```
array([4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,
       4, 4, 4, 4], dtype=int64)
```

[ ]:
```
# How to work with Filling
```

[46]:
```
df.head(15)
```

[46]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
2	NaN	NaN	NaN	NaN
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	NaN
9	4.9	3.1	1.5	setosa
10	NaN	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
14	NaN	NaN	NaN	setosa

[47]:
```
df.isna().sum()
```

[47]:
```
Sepal.Length    5
Sepal.Width     7
Petal.Length    2
Species         5
dtype: int64
```

[51]:
```
df1=df.fillna(0)            # replace the missing data with 0
df1.head(15)
```

[51]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
2	0.0	0.0	0.0	0
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	0
9	4.9	3.1	1.5	setosa
10	0.0	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
14	0.0	0.0	0.0	setosa

[52]:
```
df2=df.fillna(10)            # replace the missing data with 10
df2.head(15)
```

[52]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
2	10.0	10.0	10.0	10
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	10
9	4.9	3.1	1.5	setosa
10	10.0	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
14	10.0	10.0	10.0	setosa

[53]:
```
df.head(15)
```

[53]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
2	NaN	NaN	NaN	NaN
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	NaN
9	4.9	3.1	1.5	setosa
10	NaN	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
14	NaN	NaN	NaN	setosa

[62]:
```
df["Sepal.Length"].fillna(0,inplace=True) #only the sepal.length column's␣
↪missing values replaced by 0
```

[64]:
```
df.head(15)
```

[64]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.5	1.4	setosa
1	4.9	3.0	1.4	setosa
2	0.0	NaN	NaN	NaN
3	4.6	3.1	1.5	setosa
4	5.0	3.6	1.4	setosa
5	5.4	3.9	1.7	setosa
6	4.6	3.4	1.4	setosa
7	5.0	3.4	1.5	setosa
8	4.4	2.9	1.4	NaN
9	4.9	3.1	1.5	setosa
10	0.0	3.7	1.5	setosa
11	4.8	3.4	1.6	setosa
12	4.8	3.0	1.4	setosa
13	4.3	3.0	1.1	setosa
14	0.0	NaN	NaN	setosa

[67]:
```
df["Species"].fillna("Unknown",inplace=True) #only the sepal.length column's␣
↪missing values replaced by "Unknown"
```

[76]:
```
df.head(15)
```

[76]:

	Sepal.Length	Sepal.Width	Petal.Length	Species
0	5.1	3.50000	1.4	setosa
1	4.9	3.00000	1.4	setosa

```
2             NaN      3.04965           NaN     NaN
3             4.6      3.10000           1.5  setosa
4             5.0      3.60000           1.4  setosa
5             5.4      3.90000           1.7  setosa
6             4.6      3.40000           1.4  setosa
7             5.0      3.40000           1.5  setosa
8             4.4      2.90000           1.4     NaN
9             4.9      3.10000           1.5  setosa
10            NaN      3.70000           1.5  setosa
11            4.8      3.40000           1.6  setosa
12            4.8      3.00000           1.4  setosa
13            4.3      3.00000           1.1  setosa
14            NaN      3.04965           NaN  setosa
```

[73]:
```
df["Sepal.Width"].fillna(df["Sepal.Width"].mean(),inplace=True) #only the sepal.
 ↪length column's missing values replaced by its mean
```

[75]:
```
df.head(15)
```

[75]:
```
    Sepal.Length  Sepal.Width  Petal.Length Species
0             5.1      3.50000           1.4  setosa
1             4.9      3.00000           1.4  setosa
2             NaN      3.04965           NaN     NaN
3             4.6      3.10000           1.5  setosa
4             5.0      3.60000           1.4  setosa
5             5.4      3.90000           1.7  setosa
6             4.6      3.40000           1.4  setosa
7             5.0      3.40000           1.5  setosa
8             4.4      2.90000           1.4     NaN
9             4.9      3.10000           1.5  setosa
10            NaN      3.70000           1.5  setosa
11            4.8      3.40000           1.6  setosa
12            4.8      3.00000           1.4  setosa
13            4.3      3.00000           1.1  setosa
14            NaN      3.04965           NaN  setosa
```

[ ]: