dfcleanup

April 22, 2022

[ ]:
```
import pandas as pd
```

[ ]:
```
df = pd.read_excel("../RSA.xlsx")
```

[ ]:
```
null_values = df.isnull().sum()#isnull detect missing values and sum is used to␣
 ↪find the total number of these missing values.
print(null_values)#Print the number of null values in the dataFrame.
```

```
Galaxy     1
RA         0
Dec        0
Type       5
BT         1
Ai        20
v          2
           2
dtype: int64
```

[ ]:
```
df
```

[ ]:
```
        Galaxy        RA       Dec       Type     BT     Ai       v
0           -1        -3        -4         -9 -12.00 -14.00   -17.0   -18.0
1         F703  15 11 00  -15 16.7   S S5 2.2 -12.41   0.34  2270.0    15.0
2         HA72  13 57 39  -45 10.6   S S5 2.5 -12.83   0.38  1456.0    50.0
3      HA85-1  05 09 25  -14 51.0   S S5 2.0 -12.69   0.35  2140.0   165.0
4      HA85-2  18 52 53  -54 36.9        E 3  12.65   0.00  2761.0   113.0
…        …       …       …        …     …    …     …    …
1243    NGC986  02 31 34  -39 15.9  SB T3 1.5  11.80   0.59  2073.0   200.0
1244    NGC991  02 33 03  -07 22.0   S T5 2.0 -12.42   0.30  1530.0    15.0
1245  NGC4517A  12 29 55   00 39.9    S 7 4.0  12.65   0.40  1521.0    28.0
1246       SMC  00 51 00   -73 06.    I 9 4.5   2.79   0.25   163.0     5.0
1247       NaN     -1950     -1950        NaN    NaN    NaN     NaN     NaN

[1248 rows x 8 columns]
```

[ ]:
```
df = df.dropna(axis=0, how='any')#Drop any row (axis=0) which has one or more␣
 ↪Null values (how= 'any').
```

```
print(df)#Print the data which has been formatted to not include anymore Null␣
 ↪values and proceed to the further steps.
```

```
        Galaxy         RA        Dec        Type      BT      Ai        v
0           -1         -3         -4          -9 -12.00 -14.00    -17.0  -18.0
1         F703  15 11 00  -15 16.7    S S5 2.2 -12.41    0.34   2270.0   15.0
2         HA72  13 57 39  -45 10.6    S S5 2.5 -12.83    0.38   1456.0   50.0
3       HA85-1  05 09 25  -14 51.0    S S5 2.0 -12.69    0.35   2140.0  165.0
4       HA85-2  18 52 53  -54 36.9         E 3  12.65    0.00   2761.0  113.0
…          …       …         …         …     …     …       …      …
1242    NGC976  02 31 11   20 45.4    S R4 1.5  13.21    0.33   4362.0   58.0
1243    NGC986  02 31 34  -39 15.9   SB T3 1.5  11.80    0.59   2073.0  200.0
1244    NGC991  02 33 03  -07 22.0    S T5 2.0 -12.42    0.30   1530.0   15.0
1245  NGC4517A  12 29 55   00 39.9     S 7 4.0  12.65    0.40   1521.0   28.0
1246       SMC  00 51 00   -73 06.     I 9 4.5   2.79    0.25    163.0    5.0

[1226 rows x 8 columns]
```

[ ]:
```
df["Type"]
```

[ ]:
```
0              -9
1        S S5 2.2
2        S S5 2.5
3        S S5 2.0
4             E 3
           …
1242     S R4 1.5
1243    SB T3 1.5
1244     S T5 2.0
1245      S 7 4.0
1246      I 9 4.5
Name: Type, Length: 1226, dtype: object
```

[ ]:
```
print(type(df["Type"][0]))
```

```
<class 'int'>
```

[ ]:
```
mainclass = []
subclass = []
for i in df["Type"]:
    pars = str(i)
    parslist = pars.split(sep=" ")
    if len(parslist)>1:
        mainclass.append(parslist[0])
        subclass.append(parslist[1])
    else:
        mainclass.append(parslist[0])
        subclass.append("irrelevant")
df['MainClass'] = mainclass
df['SubClass'] = subclass
df
```

```
C:\Users\rahul\AppData\Local\Temp\ipykernel_9368\756383458.py:12:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['MainClass'] = mainclass
C:\Users\rahul\AppData\Local\Temp\ipykernel_9368\756383458.py:13:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['SubClass'] = subclass
```

```
[ ]:         Galaxy        RA       Dec       Type     BT     Ai       v         \
     0           -1        -3        -4         -9 -12.00 -14.00   -17.0  -18.0
     1         F703  15 11 00  -15 16.7   S S5 2.2 -12.41   0.34  2270.0   15.0
     2         HA72  13 57 39  -45 10.6   S S5 2.5 -12.83   0.38  1456.0   50.0
     3       HA85-1  05 09 25  -14 51.0   S S5 2.0 -12.69   0.35  2140.0  165.0
     4       HA85-2  18 52 53  -54 36.9        E 3  12.65   0.00  2761.0  113.0
     …        …       …       …        …    …    …     …    …
     1242    NGC976  02 31 11   20 45.4   S R4 1.5  13.21   0.33  4362.0   58.0
     1243    NGC986  02 31 34  -39 15.9  SB T3 1.5  11.80   0.59  2073.0  200.0
     1244    NGC991  02 33 03  -07 22.0   S T5 2.0 -12.42   0.30  1530.0   15.0
     1245  NGC4517A  12 29 55   00 39.9    S 7 4.0  12.65   0.40  1521.0   28.0
     1246       SMC  00 51 00   -73 06.    I 9 4.5   2.79   0.25   163.0    5.0

          MainClass    SubClass
     0           -9  irrelevant
     1            S          S5
     2            S          S5
     3            S          S5
     4            E           3
     …        …         …
     1242         S          R4
     1243        SB          T3
     1244         S          T5
     1245         S           7
     1246         I           9
```

```
[1226 rows x 10 columns]
```

[ ]:

```
df.to_csv("../ProcessedRSA.csv")
```

[ ]:

```
Class = []
ClassInt = []
for i in df["MainClass"]:
    pars = str(i)
    if(pars[0].upper() in ['-','1','2','3','4','5','6','7','8','9']):
        Class.append("Irrelevant")
        ClassInt.append(-1)
    elif(pars[0].upper()=="S"):
        Class.append("Spiral")
        ClassInt.append(0)
    elif(pars[0].upper()=="E"):
        Class.append("Elliptical")
        ClassInt.append(1)
    else:
        Class.append("Irregular")
        ClassInt.append(2)
Class
ClassInt
```

[ ]:

```
[-1,
 0,
 0,
 0,
 1,
 1,
 0,
 0,
 0,
 0,
 0,
 1,
 2,
 0,
 0,
 0,
 0,
 0,
 0,
 1,
 1,
1,
2,
0,
0,
2,
0,
0,
2,
0,
2,
0,
0,
0,
0,
1,
0,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
2,
0,
0,
1,
2,
1,
0,
0,
0,
0,
1,
2,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
1,
0,
0,
2,
0,
2,
0,
0,
0,
0,
0,
2,
0,
1,
0,
1,
0,
2,
0,
0,
2,
2,
1,
0,
1,
2,
1,
0,
1,
2,
0,
0,
0,
0,
1,
1,
0,
0,
1,
2,
0,
0,
1,
2,
1,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
2,
0,
0,
0,
1,
0,
0,
1,
2,
0,
0,
0,
0,
2,
2,
2,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
1,
0,
0,
0,
0,
0,
0,
0,
0,
1,
2,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
1,
0,
0,
0,
1,
2,
1,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
1,
0,
0,
0,
2,
1,
0,
0,
0,
0,
2,
2,
0,
0,
1,
0,
0,
2,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
0,
2,
1,
2,
1,
0,
0,
0,
0,
2,
2,
1,
0,
0,
2,
0,
2,
0,
0,
1,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
2,
0,
1,
0,
0,
1,
0,
0,
1,
0,
2,
0,
2,
0,
1,
0,
0,
1,
1,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
2,
0,
2,
0,
0,
2,
1,
0,
0,
1,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
2,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
1,
1,
2,
0,
2,
0,
0,
0,
2,
2,
1,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
1,
0,
0,
0,
2,
0,
0,
1,
2,
1,
1,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
1,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
1,
0,
0,
0,
0,
2,
0,
0,
0,
1,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
0,
2,
1,
0,
0,
0,
0,
0,
0,
0,
2,
2,
0,
1,
2,
0,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
2,
0,
2,
0,
0,
2,
1,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
2,
0,
0,
0,
0,
2,
0,
0,
2,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
1,
2,
2,
2,
0,
0,
1,
0,
2,
1,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
2,
1,
2,
0,
2,
1,
0,
2,
1,
1,
2,
0,
2,
0,
2,
2,
0,
2,
0,
0,
1,
0,
0,
2,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
2,
0,
2,
0,
0,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
2,
1,
2,
1,
0,
1,
2,
0,
1,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
2,
2,
0,
0,
1,
0,
0,
0,
0,
2,
0,
2,
0,
0,
0,
1,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
1,
1,
0,
0,
0,
0,
2,
2,
0,
0,
1,
0,
2,
0,
0,
0,
2,
0,
1,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
1,
1,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
1,
0,
0,
2,
2,
1,
1,
2,
0,
0,
0,
0,
0,
0,
0,
1,
1,
1,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
2,
1,
0,
0,
0,
0,
0,
0,
2,
1,
0,
2,
1,
0,
0,
0,
0,
0,
2,
0,
2,
0,
0,
0,
0,
0,
1,
0,
0,
1,
1,
0,
0,
1,
0,
0,
1,
0,
0,
0,
1,
2,
0,
2,
0,
1,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
0,
2,
0,
1,
0,
0,
0,
0,
0,
0,
0,
2,
0,
0,
0,
0,
2,
0,
1,
0,
1,
1,
0,
0,
0,
0,
2,
1,
0,
2,
0,
2,
0,
0,
0,
2,
0,
0,
0,
0,
2,
0,
0,
0,
0,
1,
0,
0,
2,
0,
2,
0,
0,
0,
2,
2,
0,
0,
0,
0,
0,
0,
0,
0,
0,
1,
0,
2,
1,
0,
0,
0,
0,
0,
0,
 0,
 0,
 0,
 0,
 0,
 1,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 1,
 0,
 0,
 0,
 2,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 0,
 2,
 0,
 1,
 0,
 1,
 1,
 2,
 …]
```

```
[ ]: df['Class'] = Class
```

```
C:\Users\rahul\AppData\Local\Temp\ipykernel_9368\4012564241.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

```
See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Class'] = Class
```

```
[ ]: df["ClassInt"] = ClassInt
```

```
C:\Users\rahul\AppData\Local\Temp\ipykernel_9368\1340064060.py:1:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df["ClassInt"] = ClassInt
```

```
[ ]: df.to_csv("../ProcessedSimplifiedRSA.csv")
```