szvz20rxp

May 7, 2023

1 Data cleaning for Algerian Forest Fire Dataset

```
[2]: import numpy as np
     import pandas as pd
     import seaborn as sns
     import matplotlib.pylab as plt
     %matplotlib inline
```

```
[3]: data = pd.read_csv(r"../Algerian_forest_fires_dataset_UPDATE.csv",header=1)
     data
```

```
[3]:     day month  year Temperature  RH  Ws Rain   FFMC  DMC    DC  ISI   BUI  \
     0    01    06  2012          29  57  18     0  65.7  3.4   7.6  1.3   3.4
     1    02    06  2012          29  61  13   1.3  64.4  4.1   7.6    1   3.9
     2    03    06  2012          26  82  22  13.1  47.1  2.5   7.1  0.3   2.7
     3    04    06  2012          25  89  13   2.5  28.6  1.3   6.9    0   1.7
     4    05    06  2012          27  77  16     0  64.8    3  14.2  1.2   3.9
     ..   ..   …   …         …  ..  ..   …   …  …   …  …   …
     241  26    09  2012          30  65  14     0  85.4   16  44.5  4.5  16.9
     242  27    09  2012          28  87  15   4.4  41.1  6.5     8  0.1   6.2
     243  28    09  2012          27  87  29   0.5  45.9  3.5   7.9  0.4   3.4
     244  29    09  2012          24  54  18   0.1  79.7  4.3  15.2  1.7   5.1
     245  30    09  2012          24  64  15   0.2  67.3  3.8  16.5  1.2   4.8

          FWI     Classes
     0    0.5   not fire
     1    0.4   not fire
     2    0.1   not fire
     3      0   not fire
     4    0.5   not fire
     ..   …           …
     241  6.5       fire
     242    0   not fire
     243  0.2   not fire
     244  0.7   not fire
     245  0.5  not fire
```

```
[246 rows x 14 columns]
```

[3]:

```
data[data.isna().any(axis=1)]
data.iloc[121:125,:]
data.drop([122,123],inplace=True)
data.reset_index(inplace=True)
data.drop(['index',"day","month","year"],axis=1,inplace=True)
data["region"] = None
data.iloc[:122,-1] = "Bejaia"
data.iloc[122:,-1] = "Abbes"
data
```

[3]:

```
    Temperature  RH  Ws Rain   FFMC  DMC    DC  ISI   BUI  FWI     Classes     \
0            29  57  18    0  65.7  3.4   7.6  1.3   3.4  0.5   not fire
1            29  61  13  1.3  64.4  4.1   7.6    1   3.9  0.4   not fire
2            26  82  22 13.1  47.1  2.5   7.1  0.3   2.7  0.1   not fire
3            25  89  13  2.5  28.6  1.3   6.9    0   1.7    0   not fire
4            27  77  16    0  64.8    3  14.2  1.2   3.9  0.5   not fire
..          …  ..  ..  …    …    …   …    …   …    …           …
239          30  65  14    0  85.4   16  44.5  4.5  16.9  6.5       fire
240          28  87  15  4.4  41.1  6.5     8  0.1   6.2    0   not fire
241          27  87  29  0.5  45.9  3.5   7.9  0.4   3.4  0.2   not fire
242          24  54  18  0.1  79.7  4.3  15.2  1.7   5.1  0.7   not fire
243          24  64  15  0.2  67.3  3.8  16.5  1.2   4.8  0.5  not fire

     region
0    Bejaia
1    Bejaia
2    Bejaia
3    Bejaia
4    Bejaia
..      …
239   Abbes
240   Abbes
241   Abbes
242   Abbes
243   Abbes

[244 rows x 12 columns]
```

2 Data cleaning operations

[4]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
```

```
Data columns (total 12 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Temperature  244 non-null    object
 1    RH          244 non-null    object
 2    Ws          244 non-null    object
 3   Rain         244 non-null    object
 4   FFMC         244 non-null    object
 5   DMC          244 non-null    object
 6   DC           244 non-null    object
 7   ISI          244 non-null    object
 8   BUI          244 non-null    object
 9   FWI          244 non-null    object
 10  Classes      244 non-null    object
 11  region       244 non-null    object
dtypes: object(12)
memory usage: 23.0+ KB
```

Getting unique values from y data column:
* Getting unique values from a column involves identifying and selecting only the distinct or unique values in that column.

```
[5]: data["Classes  "].unique()
```

```
[5]: array(['not fire   ', 'fire   ', 'fire', 'fire ', 'not fire', 'not fire ',
            'not fire     ', 'not fire    '], dtype=object)
```

Apply `str.strip()` to clean the data:
* As we can see y data has some blank spaces so we need to remove then before use.
* I have used the `.strip()` method in Python to remove the leading and trailing spaces from the data in a column.

```
[6]: data["Classes  "] = data["Classes  "].str.strip()
```

```
[7]: data
```

```
[7]:     Temperature  RH  Ws Rain  FFMC  DMC    DC  ISI   BUI  FWI Classes    \
0             29  57  18     0  65.7  3.4   7.6  1.3   3.4  0.5  not fire
1             29  61  13   1.3  64.4  4.1   7.6    1   3.9  0.4  not fire
2             26  82  22  13.1  47.1  2.5   7.1  0.3   2.7  0.1  not fire
3             25  89  13   2.5  28.6  1.3   6.9    0   1.7    0  not fire
4             27  77  16     0  64.8    3  14.2  1.2   3.9  0.5  not fire
..           …  ..  ..   …    …    …    …    …    …    …         …
239           30  65  14     0  85.4   16  44.5  4.5  16.9  6.5      fire
240           28  87  15   4.4  41.1  6.5     8  0.1   6.2    0  not fire
241           27  87  29   0.5  45.9  3.5   7.9  0.4   3.4  0.2  not fire
242           24  54  18   0.1  79.7  4.3  15.2  1.7   5.1  0.7  not fire
243           24  64  15   0.2  67.3  3.8  16.5  1.2   4.8  0.5  not fire
```

```
     region
0    Bejaia
1    Bejaia
2    Bejaia
3    Bejaia
4    Bejaia
..      …
239   Abbes
240   Abbes
241   Abbes
242   Abbes
243   Abbes

[244 rows x 12 columns]
```

[8]:
```
data["Classes  "].unique()
```

[8]:
```
array(['not fire', 'fire'], dtype=object)
```

Convert data type of all data column:
* In below code I am selecting all data which are intiger and making the column data type as float64

[9]:
```
columns = data.columns[:-2]
for i in columns:
    data[i] = data[i].astype("float64")
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 244 entries, 0 to 243
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Temperature  244 non-null    float64
 1    RH          244 non-null    float64
 2    Ws          244 non-null    float64
 3   Rain         244 non-null    float64
 4   FFMC         244 non-null    float64
 5   DMC          244 non-null    float64
 6   DC           244 non-null    float64
 7   ISI          244 non-null    float64
 8   BUI          244 non-null    float64
 9   FWI          244 non-null    float64
 10  Classes      244 non-null    object
 11  region       244 non-null    object
dtypes: float64(10), object(2)
memory usage: 23.0+ KB
```