Tutorial for Using Pandas

August 10, 2020

This is a short introduction to pandas and numpy libraries, geared mainly for new users

The objective includes: - Creat a DataFrame - Viewing a DataFrame - Selection sub-data from a DataFrame

[1]:
```
#Import the library to use
import pandas as pd
import numpy as np
```

A DataFrame: two-dimensional tabular data structure with labeled axes (rows and columns)

[2]:
```
df = pd.DataFrame({'A': 1.,
                   'B': pd.Timestamp('20201008'),
                   'C': pd.Series(1, index=list(range(4)),␣
 ↪dtype='float32'),
                   'D': np.array([3] * 4, dtype='int32'),
                   'E': ["test", "train", "test", "train"],
                   'F': 'foo'})
df
```

[2]:

	A	B	C	D	E	F
0	1.0	2020-10-08	1.0	3	test	foo
1	1.0	2020-10-08	1.0	3	train	foo
2	1.0	2020-10-08	1.0	3	test	foo
3	1.0	2020-10-08	1.0	3	train	foo

A DataFrame has columns of different types

[3]:
```
df.dtypes
```

[3]:
```
A           float64
B    datetime64[ns]
C           float32
D             int32
E            object
F            object
dtype: object
```

Convert a list to a DataFrame

[4]:
```
height_weight_list = [['David', 175, 71], ['Peter', 170, 58], ['Mark', 186, 92]]

df_index = pd.DataFrame(height_weight_list, columns=['Name', 'Height',␣
↪'Weight'])
print(df_index)
```

```
    Name  Height  Weight
0  David     175      71
1  Peter     170      58
2   Mark     186      92
```

Viewing data from the Wine Quality dataset

[5]:
```
#Read the red wine quality dataset
wine_red_dataset = pd.read_csv("winequality-red.csv", sep=';')
```

[6]:
```
wine_red_dataset
```

[6]:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	\
0	7.4	0.700	0.00	1.9	0.076	
1	7.8	0.880	0.00	2.6	0.098	
2	7.8	0.760	0.04	2.3	0.092	
3	11.2	0.280	0.56	1.9	0.075	
4	7.4	0.700	0.00	1.9	0.076	
…	…	…	…	…	…	
1594	6.2	0.600	0.08	2.0	0.090	
1595	5.9	0.550	0.10	2.2	0.062	
1596	6.3	0.510	0.13	2.3	0.076	
1597	5.9	0.645	0.12	2.0	0.075	
1598	6.0	0.310	0.47	3.6	0.067	

	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	\
0	11.0	34.0	0.99780	3.51	0.56	
1	25.0	67.0	0.99680	3.20	0.68	
2	15.0	54.0	0.99700	3.26	0.65	
3	17.0	60.0	0.99800	3.16	0.58	
4	11.0	34.0	0.99780	3.51	0.56	
…	…	…	…	…	…	
1594	32.0	44.0	0.99490	3.45	0.58	
1595	39.0	51.0	0.99512	3.52	0.76	
1596	29.0	40.0	0.99574	3.42	0.75	
1597	32.0	44.0	0.99547	3.57	0.71	
1598	18.0	42.0	0.99549	3.39	0.66	

	alcohol	quality
0	9.4	5
1	9.8	5
2	9.8	5

```
3          9.8        6
4          9.4        5
…          …          …
1594      10.5        5
1595      11.2        6
1596      11.0        6
1597      10.2        5
1598      11.0        6

[1599 rows x 12 columns]
```

Read the top rows of the dataframe

[7]:
```
wine_red_dataset.head(10)
```

[7]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076
5            7.4              0.66         0.00             1.8      0.075
6            7.9              0.60         0.06             1.6      0.069
7            7.3              0.65         0.00             1.2      0.065
8            7.8              0.58         0.02             2.0      0.073
9            7.5              0.50         0.36             6.1      0.071

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56
5                 13.0                  40.0   0.9978  3.51       0.56
6                 15.0                  59.0   0.9964  3.30       0.46
7                 15.0                  21.0   0.9946  3.39       0.47
8                  9.0                  18.0   0.9968  3.36       0.57
9                 17.0                 102.0   0.9978  3.35       0.80

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
5      9.4        5
6      9.4        5
```

```
7     10.0        7
8      9.5        7
9     10.5        5
```

Get the headers of a Dataframe

[8]:
```
list(wine_red_dataset.columns)
```

[8]:
```
['fixed acidity',
 'volatile acidity',
 'citric acid',
 'residual sugar',
 'chlorides',
 'free sulfur dioxide',
 'total sulfur dioxide',
 'density',
 'pH',
 'sulphates',
 'alcohol',
 'quality']
```

Sorting by values

[9]:
```
wine_red_dataset.sort_values(by = "fixed acidity")
```

[9]:
```
     fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
45             4.6             0.520         0.15             2.1      0.054
95             4.7             0.600         0.17             2.3      0.058
821            4.9             0.420         0.00             2.1      0.048
588            5.0             0.420         0.24             2.0      0.060
94             5.0             1.020         0.04             1.4      0.045
..             …               …             …                …        …
555           15.5             0.645         0.49             4.2      0.095
554           15.5             0.645         0.49             4.2      0.095
442           15.6             0.685         0.76             3.7      0.100
557           15.6             0.645         0.49             4.2      0.095
652           15.9             0.360         0.65             7.5      0.096

     free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
45                   8.0                  65.0  0.99340  3.90       0.56
95                  17.0                 106.0  0.99320  3.85       0.60
821                 16.0                  42.0  0.99154  3.71       0.74
588                 19.0                  50.0  0.99170  3.72       0.74
94                  41.0                  85.0  0.99380  3.75       0.48
..                   …                    …     …    …          …
555                 10.0                  23.0  1.00315  2.92       0.74
554                 10.0                  23.0  1.00315  2.92       0.74
442                  6.0                  43.0  1.00320  2.95       0.68
```

```
557                 10.0                 23.0  1.00315  2.92       0.74
652                 22.0                 71.0  0.99760  2.98       0.84

     alcohol  quality
45      13.1        4
95      12.9        6
821     14.0        7
588     14.0        8
94      10.5        4
..       …      …
555     11.1        5
554     11.1        5
442     11.2        7
557     11.1        5
652     14.9        5

[1599 rows x 12 columns]
```

Getting values of columns

[10]:
```
wine_red_dataset[['fixed acidity', 'volatile acidity', 'alcohol']]
```

[10]:
```
      fixed acidity  volatile acidity  alcohol
0               7.4             0.700      9.4
1               7.8             0.880      9.8
2               7.8             0.760      9.8
3              11.2             0.280      9.8
4               7.4             0.700      9.4
…               …                 …        …
1594            6.2             0.600     10.5
1595            5.9             0.550     11.2
1596            6.3             0.510     11.0
1597            5.9             0.645     10.2
1598            6.0             0.310     11.0

[1599 rows x 3 columns]
```

[11]:
```
#drop a column
wine_red_dataset.drop(['pH', 'quality'], axis = 1)
```

[11]:
```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0               7.4             0.700         0.00             1.9      0.076
1               7.8             0.880         0.00             2.6      0.098
2               7.8             0.760         0.04             2.3      0.092
3              11.2             0.280         0.56             1.9      0.075
4               7.4             0.700         0.00             1.9      0.076
…               …                 …            …               …        …
```

```
1594            6.2             0.600         0.08            2.0      0.090
1595            5.9             0.550         0.10            2.2      0.062
1596            6.3             0.510         0.13            2.3      0.076
1597            5.9             0.645         0.12            2.0      0.075
1598            6.0             0.310         0.47            3.6      0.067

      free sulfur dioxide  total sulfur dioxide  density  sulphates  alcohol
0                    11.0                  34.0  0.99780       0.56      9.4
1                    25.0                  67.0  0.99680       0.68      9.8
2                    15.0                  54.0  0.99700       0.65      9.8
3                    17.0                  60.0  0.99800       0.58      9.8
4                    11.0                  34.0  0.99780       0.56      9.4
…                     …                     …       …          …         …
1594                 32.0                  44.0  0.99490       0.58     10.5
1595                 39.0                  51.0  0.99512       0.76     11.2
1596                 29.0                  40.0  0.99574       0.75     11.0
1597                 32.0                  44.0  0.99547       0.71     10.2
1598                 18.0                  42.0  0.99549       0.66     11.0

[1599 rows x 10 columns]
```

Concatenate two DataFrames

[12]:
```
#Read a white wine data
wine_white_dataset = pd.read_csv("winequality-white.csv", sep=';')
```

[13]:
```
#show the shape of a DataFrame
wine_white_dataset.shape
```

[13]:
```
(4898, 12)
```

[14]:
```
wine_red_dataset.shape
```

[14]:
```
(1599, 12)
```

[15]:
```
#Concatenate to build a wine dataset (no indexes repeated)
wine_dataset = pd.concat([wine_red_dataset, wine_white_dataset],␣
 ↪ignore_index=True)
```

[16]:
```
#statistic summary of wine data:
wine_dataset.dtypes
```

[16]:
```
fixed acidity           float64
volatile acidity        float64
citric acid             float64
residual sugar          float64
chlorides               float64
```

```
free sulfur dioxide     float64
total sulfur dioxide    float64
density                 float64
pH                      float64
sulphates               float64
alcohol                 float64
quality                   int64
dtype: object
```

[17]:
```
wine_dataset.shape
```

[17]:
```
(6497, 12)
```

Selection by row index

[18]:
```
#get first 5 rows
wine_dataset[0:5]
```

[18]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0            7.4              0.70         0.00             1.9      0.076
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                 11.0                  34.0   0.9978  3.51       0.56
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality
0      9.4        5
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

Selection by labels

[19]:
```
#get rows 5 to 15 of two columns alcohol and quality
wine_dataset.loc[5:15, ['alcohol', 'quality']]
```

[19]:
```
    alcohol  quality
5       9.4        5
6       9.4        5
7      10.0        7
```

```
8       9.5        7
9      10.5        5
10      9.2        5
11     10.5        5
12      9.9        5
13      9.1        5
14      9.2        5
15      9.2        5
```

Selection by position

[20]:
```
#Get rows 1 to 4 of the wine dataset
wine_dataset.iloc[1:5]
```

[20]:
```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
1            7.8              0.88         0.00             2.6      0.098
2            7.8              0.76         0.04             2.3      0.092
3           11.2              0.28         0.56             1.9      0.075
4            7.4              0.70         0.00             1.9      0.076

   free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
1                 25.0                  67.0   0.9968  3.20       0.68
2                 15.0                  54.0   0.9970  3.26       0.65
3                 17.0                  60.0   0.9980  3.16       0.58
4                 11.0                  34.0   0.9978  3.51       0.56

   alcohol  quality
1      9.8        5
2      9.8        5
3      9.8        6
4      9.4        5
```

[21]:
```
#Getting rows 1 to 4 from columns 2 to 4
wine_dataset.iloc[1:5, 2:5]
```

[21]:
```
   citric acid  residual sugar  chlorides
1         0.00             2.6      0.098
2         0.04             2.3      0.092
3         0.56             1.9      0.075
4         0.00             1.9      0.076
```

Check for missing values

[22]:
```
wine_dataset.isnull().sum()
```

[22]:
```
fixed acidity           0
volatile acidity        0
citric acid             0
```

```
residual sugar          0
chlorides               0
free sulfur dioxide     0
total sulfur dioxide    0
density                 0
pH                      0
sulphates               0
alcohol                 0
quality                 0
dtype: int64
```

Getting all values of a column

[23]:
```
wine_dataset['quality'].value_counts()
```

[23]:
```
6    2836
5    2138
7    1079
4     216
8     193
3      30
9       5
Name: quality, dtype: int64
```

Adding headers if neccessary

[24]:
```
iris_dataset = pd.read_csv("iris.data", sep=',', header = None)
```

[25]:
```
iris_dataset
```

[25]:

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

[150 rows x 5 columns]

[26]:
```
#Adding a header
new_iris_dataset = pd.DataFrame(iris_dataset.values, columns = ["sepal_length",␣
 ↪"sepal_width", "petal_length",
```

```
                                                "petal_width", "species"])
```

[27]:
```
new_iris_dataset
```

[27]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5	3.6	1.4	0.2	Iris-setosa
..	…	…	…	…	…
145	6.7	3	5.2	2.3	Iris-virginica
146	6.3	2.5	5	1.9	Iris-virginica
147	6.5	3	5.2	2	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

Adding a header when reading a data

[28]:
```
iris_dataset_with_header = pd.read_csv("iris.data", sep=',',␣
↪names=["sepal_length", "sepal_width", "petal_length", "petal_width",␣
↪"species"])
```

[29]:
```
iris_dataset_with_header
```

[29]:

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

[150 rows x 5 columns]

[ ]:
```

```