Pandas

December 16, 2020

1 Pandas

A library to work on tabular data. To store data in the form of table.

1.0.1 Install

[1]:
```
!pip install pandas
```

```
Requirement already satisfied: pandas in c:\programdata\anaconda3\lib\site-
packages (0.25.1)
Requirement already satisfied: numpy>=1.13.3 in
c:\programdata\anaconda3\lib\site-packages (from pandas) (1.16.5)
Requirement already satisfied: python-dateutil>=2.6.1 in
c:\programdata\anaconda3\lib\site-packages (from pandas) (2.8.0)
Requirement already satisfied: pytz>=2017.2 in
c:\programdata\anaconda3\lib\site-packages (from pandas) (2019.3)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\site-
packages (from python-dateutil>=2.6.1->pandas) (1.12.0)
```

[2]:
```
import numpy as np
import pandas as pd
```

Create a data frame

[3]:
```
user_data ={
    "MarksA":np.random.randint(10,100,5),
    "MarksB":np.random.randint(10,100,5),
    "MarksC":np.random.randint(10,100,5)
}

print(user_data)
```

```
{'MarksA': array([63, 23, 64, 73, 87]), 'MarksB': array([94, 83, 61, 22, 31]),
'MarksC': array([91, 76, 54, 98, 77])}
```

[4]:
```
np.random.randint(10,100,5)
```

[4]:
```
array([18, 88, 34, 11, 85])
```

1.1 Suppose i want to create table

[5]:
```
df = pd.DataFrame(user_data)
print(df)
```

```
   MarksA  MarksB  MarksC
0      63      94      91
1      23      83      76
2      64      61      54
3      73      22      98
4      87      31      77
```

[6]:
```
#TO display in better way
df.head()
```

[6]:
```
   MarksA  MarksB  MarksC
0      63      94      91
1      23      83      76
2      64      61      54
3      73      22      98
4      87      31      77
```

[7]:
```
#Display particular n
df.head(n=3)
```

[7]:
```
   MarksA  MarksB  MarksC
0      63      94      91
1      23      83      76
2      64      61      54
```

[8]:
```
#To get the columns
df.columns
```

[8]:
```
Index(['MarksA', 'MarksB', 'MarksC'], dtype='object')
```

1.1.1 To Export Data Into CSV Files

[9]:
```
#To export the csv file
#df.to_csv('filename.csv')
df.to_csv('marks.csv')
```

[10]:
```
my_data = pd.read_csv('marks.csv')
```

[11]:
```
print(my_data)
```

```
   Unnamed: 0  MarksA  MarksB  MarksC
0           0      63      94      91
1           1      23      83      76
```

```
2           2      64      61      54
3           3      73      22      98
4           4      87      31      77
```

[12]:
```
#i dont need the first columns
my_data = my_data.drop(columns=['Unnamed: 0'] )
print(my_data)
```

```
   MarksA  MarksB  MarksC
0      63      94      91
1      23      83      76
2      64      61      54
3      73      22      98
4      87      31      77
```

1.1.2 To get Statistic of Data

[13]:
```
my_data.describe()
```

[13]:
```
           MarksA   MarksB     MarksC
count   5.000000   5.0000   5.000000
mean   62.000000  58.2000  79.200000
std    23.832751  31.4436  16.902663
min    23.000000  22.0000  54.000000
25%    63.000000  31.0000  76.000000
50%    64.000000  61.0000  77.000000
75%    73.000000  83.0000  91.000000
max    87.000000  94.0000  98.000000
```

[34]:
```
#Last 5 rows
my_data.tail()
```

[34]:
```
   MarksA  MarksB  MarksC
0      63      94      91
1      23      83      76
2      64      61      54
3      73      22      98
4      87      31      77
```

[15]:
```
#To get particular Row
df.iloc[3]
```

[15]:
```
MarksA    73
MarksB    22
MarksC    98
Name: 3, dtype: int32
```

[16]:
```
#To get particular row & col
df.iloc[3,1]
```

[16]: 22

[17]:
```
df.iloc[3][1]
```

[17]: 22

[18]:
```
#to get index of columns
idx = [df.columns.get_loc('MarksB'),df.columns.get_loc('MarksC')]
print(idx)
# In 3rd row for 1st and 2nd col
df.iloc[3,idx]
```

```
[1, 2]
```

[18]:
```
MarksB    22
MarksC    98
Name: 3, dtype: int32
```

[19]:
```
#take the first 3 rows and col 1&2
df.iloc[:3,idx]
```

[19]:

	MarksB	MarksC
0	94	91
1	83	76
2	61	54

[20]:
```
df.iloc[:3,[1,2]]
```

[20]:

	MarksB	MarksC
0	94	91
1	83	76
2	61	54

[21]:
```
##Sort your data frame on basis of marks

my_data.sort_values(by=["MarksA"],ascending=False)
```

[21]:

	MarksA	MarksB	MarksC
4	87	31	77
3	73	22	98
2	64	61	54
0	63	94	91
1	23	83	76

[22]:
```
my_data.sort_values(by=["MarksA"],ascending=True)
```

[22]:

	MarksA	MarksB	MarksC
1	23	83	76
0	63	94	91
2	64	61	54
3	73	22	98
4	87	31	77

[23]:
```
#One with highest marks in c then in A
my_data.sort_values(by=["MarksC","MarksA"],ascending=False)
```

[23]:

	MarksA	MarksB	MarksC
3	73	22	98
0	63	94	91
4	87	31	77
1	23	83	76
2	64	61	54

1.2 Pandas Into Numpy Arrays

[24]:
```
data_array = my_data.values
```

[25]:
```
print(type(my_data))
print(my_data.shape)
```

```
<class 'pandas.core.frame.DataFrame'>
(5, 3)
```

[26]:
```
print(data_array)
print(type(data_array))
print(data_array.shape)
```

```
[[63 94 91]
 [23 83 76]
 [64 61 54]
 [73 22 98]
 [87 31 77]]
<class 'numpy.ndarray'>
(5, 3)
```

1.3 Numpy Arrays Back Into Data Frame

[27]:
```
new_df = pd.
 ↪DataFrame(data_array,dtype='int32',columns=["Physics","Chemistry","Maths"])
```

[28]:
```
print(new_df)
```

```
   Physics  Chemistry  Maths
0       63         94     91
1       23         83     76
2       64         61     54
3       73         22     98
4       87         31     77
```

[29]:
```
new_df.to_csv("PCM.csv",index=False)
```

[30]:
```
#to read documentation
#new_df.to_csv?
```

[31]:
```
pcm = pd.read_csv('PCM.csv')
```

[32]:
```
print(pcm)
```

```
   Physics  Chemistry  Maths
0       63         94     91
1       23         83     76
2       64         61     54
3       73         22     98
4       87         31     77
```

[33]:
```
import pandas as pd
import matplotlib.pyplot as plt

# create 2D array of table given above
data = [['E001', 'M', 34, 123, 'Normal', 350],
        ['E002', 'F', 40, 114, 'Overweight', 450],
        ['E003', 'F', 37, 135, 'Obesity', 169],
        ['E004', 'M', 30, 139, 'Underweight', 189],
        ['E005', 'F', 44, 117, 'Underweight', 183],
        ['E006', 'M', 36, 121, 'Normal', 80],
        ['E007', 'M', 32, 133, 'Obesity', 166],
        ['E008', 'F', 26, 140, 'Normal', 120],
        ['E009', 'M', 32, 133, 'Normal', 75],
        ['E010', 'M', 36, 133, 'Underweight', 40] ]

# dataframe created with
# the above data array
df = pd.DataFrame(data, columns = ['EMPID', 'Gender',
                                   'Age', 'Sales',
                                   'BMI', 'Income'] )
print(df)

# create histogram for numeric data
df.hist()
```

```
# show plot
plt.show()
```

```
  EMPID Gender  Age  Sales          BMI  Income
0  E001      M   34    123       Normal     350
1  E002      F   40    114   Overweight     450
2  E003      F   37    135      Obesity     169
3  E004      M   30    139  Underweight     189
4  E005      F   44    117  Underweight     183
5  E006      M   36    121       Normal      80
6  E007      M   32    133      Obesity     166
7  E008      F   26    140       Normal     120
8  E009      M   32    133       Normal      75
9  E010      M   36    133  Underweight      40
```

```
<Figure size 640x480 with 4 Axes>
```

[ ]: