Pandas in Python-Day11

January 12, 2022

1 Pandas

1.1 Object Creation

[ ]:
```
import pandas as pd
import numpy as np
s= pd.Series([1,3,np.nan,5,7,8,9])
s
```

[ ]:
```
0    1.0
1    3.0
2    NaN
3    5.0
4    7.0
5    8.0
6    9.0
dtype: float64
```

1.2 Dates

[ ]:
```
dates= pd.date_range("20130101", periods=11)
```

[ ]:
```
dates
```

[ ]:
```
DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04',
               '2013-01-05', '2013-01-06', '2013-01-07', '2013-01-08',
               '2013-01-09', '2013-01-10', '2013-01-11'],
              dtype='datetime64[ns]', freq='D')
```

2 Creating a Dictionary

[ ]:
```
my_data= pd.DataFrame({ "A": 15.0, "B": pd.Timestamp("20130111"), "C": pd.
↪Series(1,index=list(range(4)), dtype="float32"), "D": np.array([3,2,1,4],␣
↪dtype="int32"), "E": pd.Categorical(["girl", "woman", "girl", "woman"]), "F":
↪ "brave"})
```

[ ]:
```
my_data
```

[ ]:
```
     A          B    C  D      E      F
0  15.0 2013-01-11  1.0  3   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data.dtypes
```

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

[ ]:
```
my_data.head()
```

[ ]:
```
     A          B    C  D      E      F
0  15.0 2013-01-11  1.0  3   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data.head(2)
```

[ ]:
```
     A          B    C  D      E      F
0  15.0 2013-01-11  1.0  3   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
```

[ ]:
```
my_data.tail(2)
```

[ ]:
```
     A          B    C  D      E      F
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data.index
```

[ ]:
```
Int64Index([0, 1, 2, 3], dtype='int64')
```

[ ]:
```
my_data.to_numpy()
```

[ ]:
```
array([[15.0, Timestamp('2013-01-11 00:00:00'), 1.0, 3, 'girl', 'brave'],
       [15.0, Timestamp('2013-01-11 00:00:00'), 1.0, 2, 'woman', 'brave'],
       [15.0, Timestamp('2013-01-11 00:00:00'), 1.0, 1, 'girl', 'brave'],
       [15.0, Timestamp('2013-01-11 00:00:00'), 1.0, 4, 'woman', 'brave']],
      dtype=object)
```

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

[ ]:
```
           A    C         D
count   4.0  4.0  4.000000
mean   15.0  1.0  2.500000
std     0.0  0.0  1.290994
min    15.0  1.0  1.000000
25%    15.0  1.0  1.750000
50%    15.0  1.0  2.500000
75%    15.0  1.0  3.250000
max    15.0  1.0  4.000000
```

[ ]:
```
my_data.T
```

[ ]:
```
                     0                   1                   2  \
A                 15.0                15.0                15.0
B  2013-01-11 00:00:00  2013-01-11 00:00:00  2013-01-11 00:00:00
C                  1.0                 1.0                 1.0
D                    3                   2                   1
E                 girl               woman                girl
F                brave               brave               brave

                     3
A                 15.0
B  2013-01-11 00:00:00
C                  1.0
D                    4
E                woman
F                brave
```

[ ]:
```
my_data
```

[ ]:
```
      A          B    C  D      E      F
0  15.0 2013-01-11  1.0  3   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data.sort_index(axis=1, ascending=False)
```

[ ]:
```
       F      E  D    C          B     A
0  brave   girl  3  1.0 2013-01-11  15.0
1  brave  woman  2  1.0 2013-01-11  15.0
2  brave   girl  1  1.0 2013-01-11  15.0
3  brave  woman  4  1.0 2013-01-11  15.0
```

[ ]:
```
my_data.sort_index(axis=0, ascending=False)
```

[ ]:
```
     A          B    C  D      E      F
3  15.0 2013-01-11  1.0  4  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
0  15.0 2013-01-11  1.0  3   girl  brave
```

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

[ ]:
```
     A          B    C  D      E      F
2  15.0 2013-01-11  1.0  1   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
0  15.0 2013-01-11  1.0  3   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data.sort_values(by="D")
```

[ ]:
```
     A          B    C  D      E      F
2  15.0 2013-01-11  1.0  1   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
0  15.0 2013-01-11  1.0  3   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data["D"]
```

[ ]:
```
0    3
1    2
2    1
3    4
Name: D, dtype: int32
```

[ ]:
```
my_data["A"]
```

[ ]:
```
0    15.0
1    15.0
2    15.0
3    15.0
Name: A, dtype: float64
```

[ ]:
```
my_data[0:4]
```

[ ]:
```
     A          B    C  D      E      F
0  15.0 2013-01-11  1.0  3   girl  brave
1  15.0 2013-01-11  1.0  2  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data[1:4]
```

[ ]:
```
     A          B    C  D      E      F
1  15.0 2013-01-11  1.0  2  woman  brave
2  15.0 2013-01-11  1.0  1   girl  brave
3  15.0 2013-01-11  1.0  4  woman  brave
```

[ ]:
```
my_data2= pd.DataFrame({ "A": 15.0, "B": pd.date_range("20130101", periods=4),␣
 ↪"C": pd.Series(1,index=list(range(4)), dtype="float32"), "D": np.
 ↪array([3,2,1,4], dtype="int32"), "E": pd.Categorical(["girl", "woman",␣
 ↪"girl", "woman"]), "F": "brave"})
```

[ ]:
```
my_data2
```

[ ]:
```
     A          B    C  D      E      F
0  15.0 2013-01-01  1.0  3   girl  brave
1  15.0 2013-01-02  1.0  2  woman  brave
2  15.0 2013-01-03  1.0  1   girl  brave
3  15.0 2013-01-04  1.0  4  woman  brave
```

[ ]:
```
date= pd.date_range("20130101", periods=4)
```

[ ]:
```
date
```

[ ]:
```
DatetimeIndex(['2013-01-01', '2013-01-02', '2013-01-03', '2013-01-04'],
dtype='datetime64[ns]', freq='D')
```

3 Slicing

[ ]:
```
my_data2.loc[:, ["A", "B"]]
#It is not working for me.
```

[ ]:
```
     A          B
0  15.0 2013-01-01
1  15.0 2013-01-02
2  15.0 2013-01-03
3  15.0 2013-01-04
```

[ ]:
```
my_data2.loc["1": "2", ["A", "B"]]
# first two terms are rows "1":"2" while the other are colums
```

[ ]:
```
     A          B
1  15.0 2013-01-02
2  15.0 2013-01-03
```

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

[ ]:
```
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0       3    male  22.0      1      0   7.2500        S  Third
```

```
1         1       1  female  38.0      1      0  71.2833        C  First
2         1       3  female  26.0      0      0   7.9250        S  Third
3         1       1  female  35.0      1      0  53.1000        S  First
4         0       3    male  35.0      0      0   8.0500        S  Third

     who  adult_male deck  embark_town alive  alone
0    man        True  NaN  Southampton    no  False
1  woman       False    C    Cherbourg   yes  False
2  woman       False  NaN  Southampton   yes   True
3  woman       False    C  Southampton   yes  False
4    man        True  NaN  Southampton    no   True
```

[ ]:
```
df.loc[["1", "4"], ["sex", "age"]]
# it is causing error because the said rows are in index
```

[ ]:
```
df2=pd.DataFrame(np.random.randn(4,4), index=date, columns=list("ABCD"))
df2
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
2013-01-03  0.017347 -0.099952  1.069820  0.263619
2013-01-04  1.148673 -0.718283  0.292902  0.587811
```

[ ]:
```
df2.loc[["20130101", "20130103"], ["A", "B"]]
```

[ ]:
```
                   A         B
2013-01-01  1.420256 -1.392784
2013-01-03  0.017347 -0.099952
```

[ ]:
```
df2.loc["20130101": "20130103", ["A", "B"]]
```

[ ]:
```
                   A         B
2013-01-01  1.420256 -1.392784
2013-01-02 -0.763163  0.200382
2013-01-03  0.017347 -0.099952
```

[ ]:
```
df2.loc["20130103", ["A", "B"]]
```

[ ]:
```
A    0.017347
B   -0.099952
Name: 2013-01-03 00:00:00, dtype: float64
```

4 Finding specific value

[ ]:
```
df2.at[dates[3],"C"]
# for row 4 and column 3 at point (3,C) rows starts from 03
# so date will be 4 0 row corresponds to data 1
```

[ ]:
```
0.2929023330957681
```

[ ]:
```
df2
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
2013-01-03  0.017347 -0.099952  1.069820  0.263619
2013-01-04  1.148673 -0.718283  0.292902  0.587811
```

5 Finding Values for Specific Rows and Columns

[ ]:
```
df2.iloc[3] # tells about the specific row
```

[ ]:
```
A    1.148673
B   -0.718283
C    0.292902
D    0.587811
Name: 2013-01-04 00:00:00, dtype: float64
```

[ ]:
```
df2.iloc[0:2] # tells about the specific row
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
```

[ ]:
```
df2.iloc[0:2, :] # tells about the specific row
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
```

[ ]:
```
df2.iloc[:, 0:2] # tells about the specific columns
```

[ ]:
```
                   A         B
2013-01-01  1.420256 -1.392784
2013-01-02 -0.763163  0.200382
2013-01-03  0.017347 -0.099952
2013-01-04  1.148673 -0.718283
```

6 Boolean Operators

[ ]:
```
df2[df2["A"]<0]
```

[ ]:
```
                   A         B         C         D
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
```

[ ]:
```
df2[df2["A"]>=0]
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-03  0.017347 -0.099952  1.069820  0.263619
2013-01-04  1.148673 -0.718283  0.292902  0.587811
```

[ ]:
```
df2[df2["A"]>=1.5] # no data
```

[ ]:
```
Empty DataFrame
Columns: [A, B, C, D]
Index: []
```

[ ]:
```
df2[df2>0]
# Nan here shows missing or the one less than zero
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256       NaN  0.760053  0.529462
2013-01-02       NaN  0.200382  0.612509       NaN
2013-01-03  0.017347       NaN  1.069820  0.263619
2013-01-04  1.148673       NaN  0.292902  0.587811
```

[ ]:
```
df3=df2.iloc[:, 0:2]
```

[ ]:
```
df3
```

[ ]:
```
                   A         B
2013-01-01  1.420256 -1.392784
2013-01-02 -0.763163  0.200382
2013-01-03  0.017347 -0.099952
2013-01-04  1.148673 -0.718283
```

[ ]:
```
df2
```

[ ]:
```
                   A         B         C         D
2013-01-01  1.420256 -1.392784  0.760053  0.529462
2013-01-02 -0.763163  0.200382  0.612509 -1.131149
2013-01-03  0.017347 -0.099952  1.069820  0.263619
2013-01-04  1.148673 -0.718283  0.292902  0.587811
```

[ ]:
```
df2["E"]=[1,2,3,4]
```

[ ]:
```
df2
```

[ ]:

	A	B	C	D	E
2013-01-01	1.420256	-1.392784	0.760053	0.529462	1
2013-01-02	-0.763163	0.200382	0.612509	-1.131149	2
2013-01-03	0.017347	-0.099952	1.069820	0.263619	3
2013-01-04	1.148673	-0.718283	0.292902	0.587811	4

[ ]:
```

```