dataframe_basics (1)

January 12, 2020

0.1 Dataframe is most commonly used object in pandas. It is a table like datastructure containing rows and columns similar to excel spreadsheet

```
[1]: import pandas as pd
     weather_data = {
         'day': ['1/1/2017','1/2/2017','1/3/2017','1/4/2017','1/5/2017','1/6/2017'],
         'temperature': [32,35,28,24,32,31],
         'windspeed': [6,7,2,7,4,2],
         'event': ['Rain', 'Sunny', 'Snow','Snow','Rain', 'Sunny']
     }
     df = pd.DataFrame(weather_data)
     #df = pd.read_csv("weather_data.csv")
     df
```

```
[1]:         day  temperature  windspeed  event
     0  1/1/2017           32          6   Rain
     1  1/2/2017           35          7  Sunny
     2  1/3/2017           28          2   Snow
     3  1/4/2017           24          7   Snow
     4  1/5/2017           32          4   Rain
     5  1/6/2017           31          2  Sunny
```

```
[2]: #df.shape #
     rows, columns = df.shape
```

```
[3]: rows
```

```
[3]: 6
```

```
[4]: df.columns
```

```
[4]: Index(['day', 'temperature', 'windspeed', 'event'], dtype='object')
```

0.2 Rows

```
[5]: df.head() # df.head(3)
```

```
[5]:          day  temperature  windspeed  event
      0  1/1/2017           32          6   Rain
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
      3  1/4/2017           24          7   Snow
      4  1/5/2017           32          4   Rain
```

```
[6]: df.head(3)
```

```
[6]:          day  temperature  windspeed  event
      0  1/1/2017           32          6   Rain
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
```

```
[7]: df.tail() # df.tail(2)
```

```
[7]:          day  temperature  windspeed  event
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
      3  1/4/2017           24          7   Snow
      4  1/5/2017           32          4   Rain
      5  1/6/2017           31          2  Sunny
```

```
[8]: df.tail(2)
```

```
[8]:          day  temperature  windspeed  event
      4  1/5/2017           32          4   Rain
      5  1/6/2017           31          2  Sunny
```

```
[9]: df[1:3]
```

```
[9]:          day  temperature  windspeed  event
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
```

```
[10]: df
```

```
[10]:         day  temperature  windspeed  event
      0  1/1/2017           32          6   Rain
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
      3  1/4/2017           24          7   Snow
      4  1/5/2017           32          4   Rain
      5  1/6/2017           31          2  Sunny
```

0.3 Columns

```
[11]: df.columns
```

```
[11]: Index(['day', 'temperature', 'windspeed', 'event'], dtype='object')
```

```
[12]: df['day'] # or df.day
```

```
[12]: 0    1/1/2017
      1    1/2/2017
      2    1/3/2017
      3    1/4/2017
      4    1/5/2017
      5    1/6/2017
      Name: day, dtype: object
```

```
[13]: df.day
```

```
[13]: 0    1/1/2017
      1    1/2/2017
      2    1/3/2017
      3    1/4/2017
      4    1/5/2017
      5    1/6/2017
      Name: day, dtype: object
```

```
[14]: type(df['day'])
```

```
[14]: pandas.core.series.Series
```

```
[15]: import pandas as pd
      df2= df[['day','temperature']]
      df2
```

```
[15]:         day  temperature
      0  1/1/2017           32
      1  1/2/2017           35
      2  1/3/2017           28
      3  1/4/2017           24
      4  1/5/2017           32
      5  1/6/2017           31
```

```
[16]: type(df[['day','temperature']])
```

```
[16]: pandas.core.frame.DataFrame
```

0.4 Operations On DataFrame

[17]:
```
df['temperature'].max()
```

[17]:
```
35
```

[18]:
```
df[df['temperature']>32]
```

[18]:
```
        day  temperature  windspeed  event
1  1/2/2017           35          7  Sunny
```

[19]:
```
df[['day','temperature']][df['temperature'] == df['temperature'].max()] # Kinda␣
 ↪doing SQL in pandas
```

[19]:
```
        day  temperature
1  1/2/2017           35
```

[20]:
```
df[df['temperature'] == df['temperature'].max()] # Kinda doing SQL in pandas
```

[20]:
```
        day  temperature  windspeed  event
1  1/2/2017           35          7  Sunny
```

[21]:
```
df['temperature'].std()
```

[21]:
```
3.8297084310253524
```

[22]:
```
df['event'].max() # But mean() won't work since data type is string
```

[22]:
```
'Sunny'
```

[23]:
```
df.describe()
```

[23]:
```
       temperature  windspeed
count     6.000000   6.000000
mean     30.333333   4.666667
std       3.829708   2.338090
min      24.000000   2.000000
25%      28.750000   2.500000
50%      31.500000   5.000000
75%      32.000000   6.750000
max      35.000000   7.000000
```

Google pandas series operations to find out list of all operations http://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.html

0.5 set_index

```
[24]: import pandas as pd

      df.set_index('day')
```

```
[24]:           temperature  windspeed  event
      day
      1/1/2017           32          6   Rain
      1/2/2017           35          7  Sunny
      1/3/2017           28          2   Snow
      1/4/2017           24          7   Snow
      1/5/2017           32          4   Rain
      1/6/2017           31          2  Sunny
```

```
[25]: df.set_index('day', inplace=True)
```

```
[26]: df
```

```
[26]:           temperature  windspeed  event
      day
      1/1/2017           32          6   Rain
      1/2/2017           35          7  Sunny
      1/3/2017           28          2   Snow
      1/4/2017           24          7   Snow
      1/5/2017           32          4   Rain
      1/6/2017           31          2  Sunny
```

```
[27]: df.index
```

```
[27]: Index(['1/1/2017', '1/2/2017', '1/3/2017', '1/4/2017', '1/5/2017', '1/6/2017'],
      dtype='object', name='day')
```

```
[28]: df.loc['1/2/2017']
```

```
[28]: temperature       35
      windspeed          7
      event          Sunny
      Name: 1/2/2017, dtype: object
```

```
[29]: df.reset_index(inplace=True)
      df.head()
```

```
[29]:         day  temperature  windspeed  event
      0  1/1/2017           32          6   Rain
      1  1/2/2017           35          7  Sunny
      2  1/3/2017           28          2   Snow
```

```
3  1/4/2017          24          7   Snow
4  1/5/2017          32          4   Rain
```

[30]:

```
df.set_index('event',inplace=True) # this is kind of building a hash map using␣
 ↪event as a key
df
```

[30]:

```
            day  temperature  windspeed
event
Rain   1/1/2017           32          6
Sunny  1/2/2017           35          7
Snow   1/3/2017           28          2
Snow   1/4/2017           24          7
Rain   1/5/2017           32          4
Sunny  1/6/2017           31          2
```

[31]:

```
df.loc['Snow']
```

[31]:

```
            day  temperature  windspeed
event
Snow   1/3/2017           28          2
Snow   1/4/2017           24          7
```