Pandas Tutorial

In [78]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
df=pd.read_csv('iris.csv')
```

In [13]:
```
df #get each row with a lable starts with 0 to end row
```

Out[13]:

	sl	sw	pl	pw	kind
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Virginica
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica

150 rows × 5 columns

In [7]:
```
print(type(df))
print(df.shape) #dimension
print(df.size)
print(df.head(2)) #1sr 2 rows
```

```
<class 'pandas.core.frame.DataFrame'>
(150, 5)
750
   sepal.length  sepal.width  petal.length  petal.width variety
0           5.1          3.5           1.4          0.2  Setosa
1           4.9          3.0           1.4          0.2  Setosa
```

In [4]:
```
df.head() #1st 5 rows
```

Out[4]:

	sepal.length	sepal.width	petal.length	petal.width	variety
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [8]:
```
df.columns
```

Out[8]:
```
Index(['sepal.length', 'sepal.width', 'petal.length', 'petal.width',
       'variety'],
      dtype='object')
```

In [9]:
```
df.columns=['sl','sw','pl','pw','kind'] #chnage column name
```

In [10]:
```
df.head()
```

Out[10]:

	sl	sw	pl	pw	kind
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [12]:
```
df.describe() #all statiscial measures
```

Out[12]:

	sl	sw	pl	pw
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.057333	3.758000	1.199333
std	0.828066	0.435866	1.765298	0.762238
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [14]:
```
df.head()
```

Out[14]:

	sl	sw	pl	pw	kind
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [17]:
```
df.iloc[2] #get index location with lable 0, ie 3rd row details
```

Out[17]:
```
sl         4.7
sw         3.2
pl         1.3
pw         0.2
kind    Setosa
Name: 2, dtype: object
```

In [18]:
```
df.loc[2]
```

Out[18]:
```
sl         4.7
sw         3.2
pl         1.3
pw         0.2
kind    Setosa
Name: 2, dtype: object
```

Deletion

In [20]:
```
df.drop(0) #delete 1st row, but actually not permanently delted
```

Out[20]:

	sl	sw	pl	pw	kind
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
5	5.4	3.9	1.7	0.4	Setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Virginica
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica

149 rows × 5 columns

In [22]:
```
df.head() #here is the 1st row not deleted
```

Out[22]:

	sl	sw	pl	pw	kind
0	5.1	3.5	1.4	0.2	Setosa
1	4.9	3.0	1.4	0.2	Setosa
2	4.7	3.2	1.3	0.2	Setosa
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa

In [28]:
```
df.drop(2, inplace=True) #no row whose label is 0 & permanently delted, by mista
```

In [29]:
```
df.head()
```

Out[29]:

	sl	sw	pl	pw	kind
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
5	5.4	3.9	1.7	0.4	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa

In [30]:
```
df.drop(5,inplace=True) #delete row with label 5
```

In [32]:
```
df.head() #no row with label 5
```

Out[32]:

	sl	sw	pl	pw	kind
3	4.6	3.1	1.5	0.2	Setosa
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa

In [33]:
```
df.index #5is deleted
```

Out[33]:
```
Index([  3,   4,   6,   7,   8,   9,  10,  11,  12,  13,
       ...
       140, 141, 142, 143, 144, 145, 146, 147, 148, 149],
      dtype='int64', length=146)
```

In [35]:
```
df.index[0] #1st row index in my data frame is 3
```

Out[35]:
```
3
```

In [36]:
```
df.index[2] #3rd row index is 6
```

Out[36]: 6

In [37]:
```
df.drop(df.index[0], inplace=True) #delete 1st row without knowing the label
```

In [38]:
```
df.head()
```

Out[38]:

	sl	sw	pl	pw	kind
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa

In [40]:
```
df[df.sw>3] #give rows where the column sw has value>3
```

Out[40]:

	sl	sw	pl	pw	kind
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa
10	5.4	3.7	1.5	0.2	Setosa
...	...	...	...	...	...
140	6.7	3.1	5.6	2.4	Virginica
141	6.9	3.1	5.1	2.3	Virginica
143	6.8	3.2	5.9	2.3	Virginica
144	6.7	3.3	5.7	2.5	Virginica
148	6.2	3.4	5.4	2.3	Virginica

63 rows × 5 columns

In [43]:
```
df[df.kind=='Setosa']
```

Out[43]:

	sl	sw	pl	pw	kind
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa
10	5.4	3.7	1.5	0.2	Setosa
11	4.8	3.4	1.6	0.2	Setosa
12	4.8	3.0	1.4	0.1	Setosa
13	4.3	3.0	1.1	0.1	Setosa
14	5.8	4.0	1.2	0.2	Setosa
15	5.7	4.4	1.5	0.4	Setosa
16	5.4	3.9	1.3	0.4	Setosa
17	5.1	3.5	1.4	0.3	Setosa
18	5.7	3.8	1.7	0.3	Setosa
19	5.1	3.8	1.5	0.3	Setosa
20	5.4	3.4	1.7	0.2	Setosa
21	5.1	3.7	1.5	0.4	Setosa
22	4.6	3.6	1.0	0.2	Setosa
23	5.1	3.3	1.7	0.5	Setosa
24	4.8	3.4	1.9	0.2	Setosa
25	5.0	3.0	1.6	0.2	Setosa
26	5.0	3.4	1.6	0.4	Setosa
27	5.2	3.5	1.5	0.2	Setosa
28	5.2	3.4	1.4	0.2	Setosa
29	4.7	3.2	1.6	0.2	Setosa
30	4.8	3.1	1.6	0.2	Setosa
31	5.4	3.4	1.5	0.4	Setosa
32	5.2	4.1	1.5	0.1	Setosa
33	5.5	4.2	1.4	0.2	Setosa
34	4.9	3.1	1.5	0.2	Setosa
35	5.0	3.2	1.2	0.2	Setosa
36	5.5	3.5	1.3	0.2	Setosa
37	4.9	3.6	1.4	0.1	Setosa

	sl	sw	pl	pw	kind
38	4.4	3.0	1.3	0.2	Setosa
39	5.1	3.4	1.5	0.2	Setosa
40	5.0	3.5	1.3	0.3	Setosa
41	4.5	2.3	1.3	0.3	Setosa
42	4.4	3.2	1.3	0.2	Setosa
43	5.0	3.5	1.6	0.6	Setosa
44	5.1	3.8	1.9	0.4	Setosa
45	4.8	3.0	1.4	0.3	Setosa
46	5.1	3.8	1.6	0.2	Setosa
47	4.6	3.2	1.4	0.2	Setosa
48	5.3	3.7	1.5	0.2	Setosa
49	5.0	3.3	1.4	0.2	Setosa

In [44]:
```
df.head()
```

Out[44]:

	sl	sw	pl	pw	kind
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa

In [46]:
```
df.iloc[0] #get the 1st index row
```

Out[46]:
```
sl         5.0
sw         3.6
pl         1.4
pw         0.2
kind    Setosa
Name: 4, dtype: object
```

In [48]:
```
df.loc[7] #get the row with location/label 7, not the 8th index row
```

Out[48]:
```
sl         5.0
sw         3.4
pl         1.5
pw         0.2
kind    Setosa
Name: 7, dtype: object
```

Add new row

In [55]:
```
df.loc[0]=[2.3,3.4,4.2,1.2,'nigo'] #intesr row where label is 0 at last
```

In [56]:
```
df.head() #not at first its added
```

Out[56]:

	sl	sw	pl	pw	kind
4	5.0	3.6	1.4	0.2	Setosa
6	4.6	3.4	1.4	0.3	Setosa
7	5.0	3.4	1.5	0.2	Setosa
8	4.4	2.9	1.4	0.2	Setosa
9	4.9	3.1	1.5	0.1	Setosa

In [57]:
```
df.tail() #added at last
```

Out[57]:

	sl	sw	pl	pw	kind
146	6.3	2.5	5.0	1.9	Virginica
147	6.5	3.0	5.2	2.0	Virginica
148	6.2	3.4	5.4	2.3	Virginica
149	5.9	3.0	5.1	1.8	Virginica
0	2.3	3.4	4.2	1.2	nigo

In [59]:
```
df.reset_index() #new column added called 'index'
```

Out[59]:

	index	sl	sw	pl	pw	kind
0	4	5.0	3.6	1.4	0.2	Setosa
1	6	4.6	3.4	1.4	0.3	Setosa
2	7	5.0	3.4	1.5	0.2	Setosa
3	8	4.4	2.9	1.4	0.2	Setosa
4	9	4.9	3.1	1.5	0.1	Setosa
...	...	...	...	...	...	...
141	146	6.3	2.5	5.0	1.9	Virginica
142	147	6.5	3.0	5.2	2.0	Virginica
143	148	6.2	3.4	5.4	2.3	Virginica
144	149	5.9	3.0	5.1	1.8	Virginica
145	0	2.3	3.4	4.2	1.2	nigo

146 rows × 6 columns

In [64]:
```
df.reset_index(drop=True, inplace=True) #delete index column
#& get the dataframe serially in correct order}
```

In [63]:
```
df.head()
```

Out[63]:

	sl	sw	pl	pw	kind
0	5.0	3.6	1.4	0.2	Setosa
1	4.6	3.4	1.4	0.3	Setosa
2	5.0	3.4	1.5	0.2	Setosa
3	4.4	2.9	1.4	0.2	Setosa
4	4.9	3.1	1.5	0.1	Setosa

df.drop('sl') #delete column but by default it checks in row & its not present so get error
TO delete column give axis=1

In [66]:
```
df.drop('sl', axis=1) #delete column with name sl, but deletion in temporary
```

Out[66]:

	sw	pl	pw	kind
0	3.6	1.4	0.2	Setosa
1	3.4	1.4	0.3	Setosa
2	3.4	1.5	0.2	Setosa
3	2.9	1.4	0.2	Setosa
4	3.1	1.5	0.1	Setosa
...	...	...	...	...
141	2.5	5.0	1.9	Virginica
142	3.0	5.2	2.0	Virginica
143	3.4	5.4	2.3	Virginica
144	3.0	5.1	1.8	Virginica
145	3.4	4.2	1.2	nigo

146 rows × 4 columns

In [67]:
```
df.head()
```

Out[67]:

	sl	sw	pl	pw	kind
0	5.0	3.6	1.4	0.2	Setosa
1	4.6	3.4	1.4	0.3	Setosa
2	5.0	3.4	1.5	0.2	Setosa
3	4.4	2.9	1.4	0.2	Setosa
4	4.9	3.1	1.5	0.1	Setosa

In [68]:
```
df.drop('sl', axis=1, inplace=True) #axis=1-->check in column &
                                    #inplace=True-->change in actual df not in c
```

In [69]:
```
df.head()
```

Out[69]:

	sw	pl	pw	kind
0	3.6	1.4	0.2	Setosa
1	3.4	1.4	0.3	Setosa
2	3.4	1.5	0.2	Setosa
3	2.9	1.4	0.2	Setosa
4	3.1	1.5	0.1	Setosa

Add New Column

In [70]:
```
df['sum']=df['sw']+df['pw']
```

In [72]:
```
df.head()
```

Out[72]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
1	3.4	1.4	0.3	Setosa	3.7
2	3.4	1.5	0.2	Setosa	3.6
3	2.9	1.4	0.2	Setosa	3.1
4	3.1	1.5	0.1	Setosa	3.2

In [73]:
```
df['new_col']=1 #new column with value=1
```

In [74]:
```
df.head()
```

Out[74]:

	sw	pl	pw	kind	sum	new_col
0	3.6	1.4	0.2	Setosa	3.8	1
1	3.4	1.4	0.3	Setosa	3.7	1
2	3.4	1.5	0.2	Setosa	3.6	1
3	2.9	1.4	0.2	Setosa	3.1	1
4	3.1	1.5	0.1	Setosa	3.2	1

In [ ]:
```
del df['new_col'] #delete it
```

In [76]:
```
df.head()
```

Out[76]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
1	3.4	1.4	0.3	Setosa	3.7
2	3.4	1.5	0.2	Setosa	3.6
3	2.9	1.4	0.2	Setosa	3.1
4	3.1	1.5	0.1	Setosa	3.2

NaN in data set ("Not a Number")

we have to delete NaN or replace nan with some value

In [79]:
```
df.iloc[1:3, 0:2]=np.nan #adding some nan value to dataset
                         #added to 1,2 row and 0,1 column position
```

In [80]:
```
df.head()
```

Out[80]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
1	NaN	NaN	0.3	Setosa	3.7
2	NaN	NaN	0.2	Setosa	3.6
3	2.9	1.4	0.2	Setosa	3.1
4	3.1	1.5	0.1	Setosa	3.2

In [81]:
```
df.describe()
```

Out[81]:

	sw	pl	pw	sum
count	144.000000	144.000000	146.000000	146.000000
mean	3.045139	3.872917	1.232192	4.282192
std	0.436424	1.724899	0.750993	0.721940
min	2.000000	1.000000	0.100000	2.600000
25%	2.800000	1.600000	0.325000	3.700000
50%	3.000000	4.400000	1.300000	4.200000
75%	3.300000	5.100000	1.800000	4.800000
max	4.400000	6.900000	2.500000	6.100000

In [83]:
```
df.dropna() #not permanently
```

Out[83]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
3	2.9	1.4	0.2	Setosa	3.1
4	3.1	1.5	0.1	Setosa	3.2
5	3.7	1.5	0.2	Setosa	3.9
6	3.4	1.6	0.2	Setosa	3.6
...	...	...	...	...	...
141	2.5	5.0	1.9	Virginica	4.4
142	3.0	5.2	2.0	Virginica	5.0
143	3.4	5.4	2.3	Virginica	5.7
144	3.0	5.1	1.8	Virginica	4.8
145	3.4	4.2	1.2	nigo	4.6

144 rows × 5 columns

In [86]:
```
df.dropna(inplace=True) #deleted permenantly & also deleted the rows having the
```

In [85]:
```
df.head()
```

Out[85]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
3	2.9	1.4	0.2	Setosa	3.1
4	3.1	1.5	0.1	Setosa	3.2
5	3.7	1.5	0.2	Setosa	3.9
6	3.4	1.6	0.2	Setosa	3.6

In [87]:
```
df.reset_index(drop=True,inplace=True)
```

In [88]:
```
df.head()
```

Out[88]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
1	2.9	1.4	0.2	Setosa	3.1
2	3.1	1.5	0.1	Setosa	3.2
3	3.7	1.5	0.2	Setosa	3.9
4	3.4	1.6	0.2	Setosa	3.6

In [89]:
```
df.iloc[1:4, 0:2]=np.nan
```

In [90]:
```
df.head(7)
```

Out[90]:

	sw	pl	pw	kind	sum
0	3.6	1.4	0.2	Setosa	3.8
1	NaN	NaN	0.2	Setosa	3.1
2	NaN	NaN	0.1	Setosa	3.2
3	NaN	NaN	0.2	Setosa	3.9
4	3.4	1.6	0.2	Setosa	3.6
5	3.0	1.4	0.1	Setosa	3.1
6	3.0	1.1	0.1	Setosa	3.1

In [95]:
```
m=df.sw.mean()
df.fillna(m) #fill the NaN value of 'ALL' column with mean value of sw
```

Out[95]:

	sw	pl	pw	kind	sum
0	3.600000	1.400000	0.2	Setosa	3.8
1	3.041135	3.041135	0.2	Setosa	3.1
2	3.041135	3.041135	0.1	Setosa	3.2
3	3.041135	3.041135	0.2	Setosa	3.9
4	3.400000	1.600000	0.2	Setosa	3.6
...	...	...	...	...	...
139	2.500000	5.000000	1.9	Virginica	4.4
140	3.000000	5.200000	2.0	Virginica	5.0
141	3.400000	5.400000	2.3	Virginica	5.7
142	3.000000	5.100000	1.8	Virginica	4.8
143	3.400000	4.200000	1.2	nigo	4.6

144 rows × 5 columns

In [96]:
```
df.sw.fillna(m,inplace=True) #ffill the NaN value of 'sw' column with mean value
```

In [98]:
```
df.head(8) #pl column has left with NaN value
```

Out[98]:

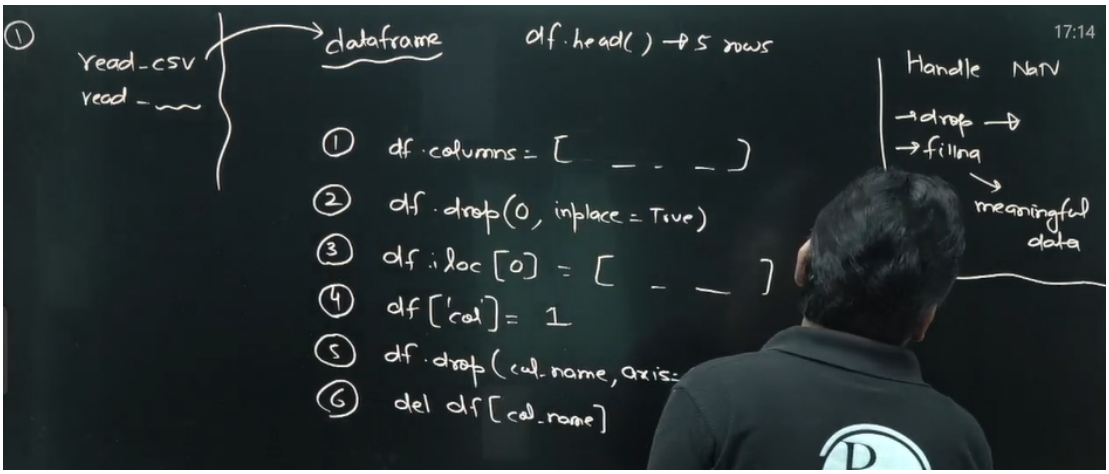
	sw	pl	pw	kind	sum
0	3.600000	1.4	0.2	Setosa	3.8
1	3.041135	NaN	0.2	Setosa	3.1
2	3.041135	NaN	0.1	Setosa	3.2
3	3.041135	NaN	0.2	Setosa	3.9
4	3.400000	1.6	0.2	Setosa	3.6
5	3.000000	1.4	0.1	Setosa	3.1
6	3.000000	1.1	0.1	Setosa	3.1
7	4.000000	1.2	0.2	Setosa	4.2

In [99]:
```
df.pl.fillna(0,inplace=True)
```

In [100…
```
df.head()
```

Out[100…

	sw	pl	pw	kind	sum
0	3.600000	1.4	0.2	Setosa	3.8
1	3.041135	0.0	0.2	Setosa	3.1
2	3.041135	0.0	0.1	Setosa	3.2
3	3.041135	0.0	0.2	Setosa	3.9
4	3.400000	1.6	0.2	Setosa	3.6



In [ ]:
```

```