Read data set

In [1]:
```
# method - 1
import pandas as pd
# read data frame
df=pd.read_csv('C:/Users/omkan/Desktop/Iris.csv')
df
```

Out[1]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

In [2]:
```
# method - 2
import pandas as pd
# read data frame
df=pd.read_csv('C:\\Users\\omkan\\Desktop\\Iris.csv')
df
```

Out[2]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

In [3]:
```
# method - 3
import pandas as pd
# read data frame
df=pd.read_csv(r'C:\Users\omkan\Desktop\Iris.csv')
# if excel-   pd.read_excel or pd. read_csv
df
```

Out[3]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

--read data set in colab (colab is google online platform for ML,AI,DS,DL)

#from google.colab import drive

#drive.mount('/content/drive/')

--manually upload file in drive then

#df=pd.read_csv('//content//drive//My Drive//DataSet.csv',encoding='latin1') #df.head(2) -

-all columns or row #pd.reset_option('max_row',None)

#pd.reset_option('max_column',None)

understanding_analysing the data Set

> Methods

- here v stands for variable
- v.head(), v.tail(), v.info(), v.describe(), v.corr()
- v.isnull().sum(), v.notnull().sum(), v.count()
- v['column_name'].value_counts().sort_value()
- v['column_name'].unique()
- v['column_name'].nunique()
- v.sample(no.)

> Attributes

- v.size
- v.shape
- v.columns
- v.ndim

```
# importent method to extract big amount of data sets
df.groupby(), df.query(), df.iloc[],df.loc[], extraction v[]_operators
```

In [ ]:

In [4]:
```
df.iloc[[98],:]
```

Out[4]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
98	99	5.1	2.5	3.0	1.1	Iris-versicolor

In [5]:
```
df['Species'].value_counts()
```

Out[5]:
```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
```

In [6]:
```
df['Species']=df['Species'].replace({'Iris-setosa':0,'Iris-virginica':1,'Iris-versicolor':2})
```

In [7]:
```
df['Species'].value_counts()
```

Out[7]:
```
0    50
2    50
1    50
Name: Species, dtype: int64
```

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

Out[8]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
145	146	6.7	3.0	5.2	2.3	1
146	147	6.3	2.5	5.0	1.9	1
147	148	6.5	3.0	5.2	2.0	1
148	149	6.2	3.4	5.4	2.3	1
149	150	5.9	3.0	5.1	1.8	1

In [9]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             150 non-null    int64
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    int64
dtypes: float64(4), int64(2)
memory usage: 7.2 KB
```

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

Out[10]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	75.500000	5.843333	3.054000	3.758667	1.198667	1.000000
std	43.445368	0.828066	0.433594	1.764420	0.763161	0.819232
min	1.000000	4.300000	2.000000	1.000000	0.100000	0.000000
25%	38.250000	5.100000	2.800000	1.600000	0.300000	0.000000
50%	75.500000	5.800000	3.000000	4.350000	1.300000	1.000000
75%	112.750000	6.400000	3.300000	5.100000	1.800000	2.000000
max	150.000000	7.900000	4.400000	6.900000	2.500000	2.000000

In [11]:
```
df.corr()
```

Out[11]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
Id	1.000000	0.716676	-0.397729	0.882747	0.899759	0.471415
SepalLengthCm	0.716676	1.000000	-0.109369	0.871754	0.817954	0.460039
SepalWidthCm	-0.397729	-0.109369	1.000000	-0.420516	-0.356544	-0.612165
PetalLengthCm	0.882747	0.871754	-0.420516	1.000000	0.962757	0.649101
PetalWidthCm	0.899759	0.817954	-0.356544	0.962757	1.000000	0.580749
Species	0.471415	0.460039	-0.612165	0.649101	0.580749	1.000000

In [12]:
```
df.isnull().sum()
```

Out[12]:
```
Id              0
SepalLengthCm   0
SepalWidthCm    0
PetalLengthCm   0
PetalWidthCm    0
Species         0
dtype: int64
```

In [13]:
```
df.count()
```

Out[13]:
```
Id              150
SepalLengthCm   150
SepalWidthCm    150
PetalLengthCm   150
PetalWidthCm    150
Species         150
dtype: int64
```

In [14]:
```
df['Species'].value_counts()
```

Out[14]:
```
0    50
2    50
1    50
Name: Species, dtype: int64
```

In [15]:
```
df['Species'].value_counts().sort_values()
```

Out[15]:
```
0    50
2    50
1    50
Name: Species, dtype: int64
```

In [16]:
```
df['Species'].sort_values()
```

Out[16]:
```
0     0
27    0
28    0
29    0
30    0
     ..
78    2
77    2
76    2
86    2
74    2
Name: Species, Length: 150, dtype: int64
```

In [17]:
```
df['Species'].unique()
```

Out[17]:
```
array([0, 2, 1], dtype=int64)
```

In [18]:
```
df['Species'].nunique()
```

Out[18]:
```
3
```

In [19]:
```
df.sample(7) # randomly provide values
```

Out[19]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
66	67	5.6	3.0	4.5	1.5	2
7	8	5.0	3.4	1.5	0.2	0
3	4	4.6	3.1	1.5	0.2	0
86	87	6.7	3.1	4.7	1.5	2
51	52	6.4	3.2	4.5	1.5	2
105	106	7.6	3.0	6.6	2.1	1
67	68	5.8	2.7	4.1	1.0	2

Attributes

In [20]:
```
df.columns
```

Out[20]:
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species'],
      dtype='object')
```

In [21]:
```
df.dtypes
```

Out[21]:
```
Id                int64
SepalLengthCm    float64
SepalWidthCm     float64
PetalLengthCm    float64
PetalWidthCm     float64
Species            int64
dtype: object
```

In [22]:
```
df.shape,df.size,type(df)
```

Out[22]:
```
((150, 6), 900, pandas.core.frame.DataFrame)
```

In [ ]:

In [23]:
```
# creat series from list
import pandas as pd
l=[1,2,3,4]
s=pd.Series(l,index=('a','b','c','d'),dtype='float')
s,type(s),s.shape,s.size
```

Out[23]:
```
(a    1.0
 b    2.0
 c    3.0
 d    4.0
 dtype: float64,
 pandas.core.series.Series,
 (4,),
 4)
```

In [ ]:

In [24]:
```
# statastical operations with Series
s.mean(),s.mode(),s.median()
```

Out[24]:
```
(2.5,
 0    1.0
 1    2.0
 2    3.0
 3    4.0
 dtype: float64,
 2.5)
```

In [25]:
```
# create csv from series
s.to_csv('series_first.csv',index=False)
```

In [26]:
```
a=[1,2,3,4,5]            # create Series from list
a
import pandas as pd
a=pd.Series(a,index=('a','b','c','d','e'),dtype='str')
a,type(a)
```

Out[26]:
```
(a    1
 b    2
 c    3
 d    4
 e    5
 dtype: object,
 pandas.core.series.Series)
```

In [27]:
```
# create series form dict
s_d={'a':1,'b':2,'c':3}
s_d=pd.Series(s_d)
s_d
```

Out[27]:
```
a    1
b    2
c    3
dtype: int64
```

In [28]:
```
# create a series from tuple
s_t=(1,2,3,4,5)
pd.Series(s_t)
```

Out[28]:
```
0    1
1    2
2    3
3    4
4    5
dtype: int64
```

DataFrame

In [29]:
```
l=[['om',100,'Indore'],        # Create from list
   ['kant',98,'Ujjain'],
   ['sharma',99,'Bhopal']]
d_l=pd.DataFrame(l,columns=('name','marks','place'),index=('r1','r2','r3'))
d_l
```

Out[29]:

	name	marks	place
r1	om	100	Indore
r2	kant	98	Ujjain
r3	sharma	99	Bhopal

In [30]:
```
import pandas as pd
df=pd.read_csv("C:\\Users\\omkan\\Desktop\\Iris.csv")
df.ndim
```

Out[30]: 2

In [31]:
```
d=pd.DataFrame({'Name':['om','kant','sharma'],
                'Marks':[100,98,99]
                })
d
```

Out[31]:

	Name	Marks
0	om	100
1	kant	98
2	sharma	99

In [32]:
```
# creating Dataframe from dictionary
d={'name':['om','kant','sharma'],'Marks':[10,20,30]}
d=pd.DataFrame(d)
d
```

Out[32]:

	name	Marks
0	om	10
1	kant	20
2	sharma	30

create csv file from dataframe

- v.to_csv('file_name.csv',index=False)

In [33]:
```
# create csv file from dataframe
d.to_csv('DataFrame_first.csv',index=False)
```

In [34]:
```
d=pd.read_csv('DataFrame_first.csv')
d
```

Out[34]:

	name	Marks
0	om	10
1	kant	20
2	sharma	30

In [35]:
```
d['name']                 # extract columns as Series
```

Out[35]:
```
0        om
1      kant
2    sharma
Name: name, dtype: object
```

In [36]:
```
d['Marks']
```

Out[36]:
```
0    10
1    20
2    30
Name: Marks, dtype: int64
```

In [37]:
```
e_a=d['name']
e_a
```

Out[37]:
```
0        om
1      kant
2    sharma
Name: name, dtype: object
```

```
# if you use duble square Bracket then it will become dic
```

In [38]:
```
#e_a[['name']]
w=[[e_a]] # it will become list
type(w)
```

Out[38]:
```
list
```

In [39]:
```
d[['name']]
```

Out[39]:

	name
0	om
1	kant
2	sharma

In [40]:
```
# extracting Series from DataFrame -
s=pd.Series(d['name'])
s1=d['name']
s,s1
```

Out[40]:
```
(0        om
 1      kant
 2    sharma
 Name: name, dtype: object,
 0        om
 1      kant
 2    sharma
 Name: name, dtype: object)
```

Casting in a list & numpy

- v.tolist() # sheries case
- v.index.tolist()
- v.values.tolist())
- v.columns.tolist()
- v.values as attribute convert array
- v.to_numpy()

In [41]:
```
# most Importent  in matplotlinb and seaborn
# value, index & columns
v_s=s.values   # casting in array
i_s=s.index
c_d=d.columns
v_s,i_s,c_d
```

Out[41]:
```
(array(['om', 'kant', 'sharma'], dtype=object),
 RangeIndex(start=0, stop=3, step=1),
 Index(['name', 'Marks'], dtype='object'))
```

In [42]:
```
pd.DataFrame(v_s)
```

Out[42]:

	0
0	om
1	kant
2	sharma

In [43]:
```
# convert value and index into list
s_v_l=s.values.tolist()  # array then list - v.values.tolist()
s_i_l=s.index.tolist()
s_v_l,s_i_l
```

Out[43]:
```
(['om', 'kant', 'sharma'], [0, 1, 2])
```

Use numpys to create DataFrame

In [44]:
```
import numpy as np
import random
arr=np.random.rand(10).reshape(2,5)
pd.DataFrame(arr)
```

Out[44]:

	0	1	2	3	4
0	0.857602	0.275393	0.24882	0.106394	0.523358
1	0.296664	0.538394	0.17877	0.190496	0.038185

In [45]:
```
# extracting elements from series  through indexing
s[2],s[0],s[1]
```

Out[45]:
```
('sharma', 'om', 'kant')
```

In [46]:
```
# extracting elements from dictionary through slicing
s[:2],s[-5:-1]
```

Out[46]:
```
(0      om
 1    kant
 Name: name, dtype: object,
 0      om
 1    kant
 Name: name, dtype: object)
```

In [47]:
```
s=pd.Series([1,2,3,4,4,6,5])
s.mean(),s.mode(),s.median(),s.count(),s.cumsum(),s.cumprod() # cumulative sum & product
```

Out[47]:
```
(3.5714285714285716,
 0    4
 dtype: int64,
 4.0,
 7,
 0     1
 1     3
 2     6
 3    10
 4    14
 5    20
 6    25
 dtype: int64,
 0       1
 1       2
 2       6
 3      24
 4      96
 5     576
 6    2880
 dtype: int64)
```

DataFrame - most importent creating dataframe from Series & add Series

In [48]:
```
# DataFrame - most importent creating dataframe from Series
l=[1,2,3,4,5]
l2=['a','b','c','d','e']
s=pd.Series(l)
s1=pd.Series(l2)
s,s1
```

Out[48]:
```
(0    1
 1    2
 2    3
 3    4
 4    5
 dtype: int64,
 0    a
 1    b
 2    c
 3    d
 4    e
 dtype: object)
```

In [49]:
```
d_s=pd.DataFrame({'int':s,'str':s1})
d_s
```

Out[49]:

	int	str
0	1	a
1	2	b
2	3	c
3	4	d
4	5	e

In [50]:
```
d_s.shape
```

Out[50]:
```
(5, 2)
```

In [51]:
```
# adding sering in dataframe
l='A','B','C','D','E'
s_a_d=pd.Series(l)
s_a_d,s_a_d.shape
```

Out[51]:
```
(0    A
 1    B
 2    C
 3    D
 4    E
 dtype: object,
 (5,))
```

In [52]:
```
d_s_a=d_s['U_str']=s_a_d   # v[key(given_name)]=value(series as value)
d_s_a # here single value assign to multiple variable
```

Out[52]:
```
0    A
1    B
2    C
3    D
4    E
dtype: object
```

In [53]:
```
l=[1.1,1.2,1.3,1.4,1.5]
f_s=pd.Series(l)
d_s['float']=f_s    # variable['key'(a key give name)]=value(pass Series)
```

In [54]:
```
d_s
```

Out[54]:

	int	str	U_str	float
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

set_index() - this function for set_index

- allow to assign a column name as a index of a dataframe

In [55]:
```
d_s.set_index('float')
```

Out[55]:

float	int	str	U_str
1.1	1	a	A
1.2	2	b	B
1.3	3	c	C
1.4	4	d	D
1.5	5	e	E

sort_index()

sort_values()

- list items
- list items

In [56]:
```
d_l.sort_index(ascending=False)
```

Out[56]:

	name	marks	place
r3	sharma	99	Bhopal
r2	kant	98	Ujjain
r1	om	100	Indore

In [57]:
```
d_l.sort_index(axis=1) # OR to print column names in reverse alphabetical order
# v.sort_index(axis=1,ascending=False)
```

Out[57]:

	marks	name	place
r1	100	om	Indore
r2	98	kant	Ujjain
r3	99	sharma	Bhopal

In [58]:
```
d_l.sort_index(axis=0,ascending=False) # alphabetical order
```

Out[58]:

	name	marks	place
r3	sharma	99	Bhopal
r2	kant	98	Ujjain
r1	om	100	Indore

In [59]:
```
d_l.sort_index(axis=0)
```

Out[59]:

	name	marks	place
r1	om	100	Indore
r2	kant	98	Ujjain
r3	sharma	99	Bhopal

In [60]:
```
#v.sort_value(by='col', ascending=fasle)  # colums wise asceding true or false
```

In [61]:
```
d=pd.Series([33,32,31],index=('r1','r2','r3'))
d_l['ages']=d
d_l
```

Out[61]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [62]:
```
d_l.sort_values(by='place', ascending=False)
```

Out[62]:

	name	marks	place	ages
r2	kant	98	Ujjain	32
r1	om	100	Indore	33
r3	sharma	99	Bhopal	31

In [63]:
```
d_l.sort_values(by=['ages','marks'],ascending=False)
```

Out[63]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [ ]:
```

```

In [64]:
```
# access columns values in series
# method  1 in Seies
d_s['str']
```

Out[64]:
```
0    a
1    b
2    c
3    d
4    e
Name: str, dtype: object
```

In [65]:
```
# method 2 in Series
d_s.str
```

Out[65]:
```
0    a
1    b
2    c
3    d
4    e
Name: str, dtype: object
```

Accessing one/multiple colums in DataFrame

- v['col_1'] extract as Series(single square brecket)
- v[['col_1']]extract as Data Frame (duble squre brecket)
- v[['col_1,col_2,col_n']]
- v.loc[[r_index_no1,r_index_no2],[col_names]]
- v.loc[slicing :,[col_names]]
- v.loc[[row_no./names], slicing :]
- v.loc[:,:]
- v.iloc[1:2 (sliceing),[cl_index_n1,cl_index_n2]]
- v.iloc[[1,2],[0]]
- v.iloc[[r1,r3(row_indexing)],[c0,c4(column_indexing)]]
- v.iloc[:,:]
- v.iloc[start:end:step(row),start:end:step(columns)]
- v.dorp() - drop(['row_no./col_name'],axis=0(by defalt row)/1(column))
- inplace=True - in any func means permanent change in present col.
- v.drop_duplicates()

In [66]:
```
# accessing multiple colums values in DataFrame
# method 1
d_s[['str','U_str']]      # use double squre Brecket
```

Out[66]:

	str	U_str
0	a	A
1	b	B
2	c	C
3	d	D
4	e	E

In [67]:
```
# loc function in row can indexing  and slicing but column write a column name
# method - 2 with loc function
d_s.loc[[1,2,4,1],['str','int']]  # v.loc[[1,3,4(random_indexing)],['cl_name1','cl_name2']]
```

Out[67]:

	str	int
1	b	2
2	c	3
4	e	5
1	b	2

In [68]:
```
# loc function in row can indexing  and slicing but column write a column name
# method - 3  with loc function
d_s.loc[1:2,['int']] # v.loc[ 2:30 (slicing),['cl_name1,cl_name2']
```

Out[68]:

	int
1	2
2	3

In [69]:
```
d_s
```

Out[69]:

	int	str	U_str	float
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

In [70]:
```
d_s.loc[:,['float']]
```

Out[70]:

	float
0	1.1
1	1.2
2	1.3
3	1.4
4	1.5

In [71]:
```
import numpy as np
a=np.array([[1,2,3],[4,5,6],[7,8,9]])
a[[1,2,0,2,1],[0,1,1,2,1]]
```

Out[71]:
```
array([4, 8, 2, 9, 5])
```

In [72]:
```
# method -1 in v.iloc[]
# v.iloc[ 1:2 (sliceing),[cl_index_n1,cl_index_n2]]
```

In [73]:
```
d_s.iloc[1:,[0,3]]      # index no. use for columns
```

Out[73]:

	int	float
1	2	1.2
2	3	1.3
3	4	1.4
4	5	1.5

In [74]:
```
d_s.iloc[[1,2],[0]]    # v.iloc[[r1,r3(row_indexing)],[c0,c4(column_indexing)]]
```

Out[74]:

	int
1	2
2	3

In [75]:
```
d_s.iloc[:,:]           # v.iloc[start:end:step(row),start:end:step(columns)]
```

Out[75]:

	int	str	U_str	float
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

In [76]:
```
df.loc[[10,130,80],['PetalLengthCm']]    # with indexing
```

Out[76]:

	PetalLengthCm
10	1.5
130	6.1
80	3.8

In [77]:
```
df.loc[:,['SepalLengthCm','PetalLengthCm']].head(2)
```

Out[77]:

	SepalLengthCm	PetalLengthCm
0	5.1	1.4
1	4.9	1.4

In [78]:
```
df.loc[(df['SepalWidthCm']>4)&(df['SepalLengthCm']>5)&(df['Species']=='Iris-setosa')]
```

Out[78]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
15	16	5.7	4.4	1.5	0.4	Iris-setosa
32	33	5.2	4.1	1.5	0.1	Iris-setosa
33	34	5.5	4.2	1.4	0.2	Iris-setosa

In [79]:
```
df.iloc[[1,125,93],[1,2,3,4,5]] # with indexing
```

Out[79]:

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	4.9	3.0	1.4	0.2	Iris-setosa
125	7.2	3.2	6.0	1.8	Iris-virginica
93	5.0	2.3	3.3	1.0	Iris-versicolor

In [80]:
```
df.iloc[3:10,1:]
```

Out[80]:

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa

In [ ]:
```
```

Filtter DataFrame

query

- v.query('col_name < No.')
- complax queries

In [81]:
```
d_s
```

Out[81]:

	int	str	U_str	float
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

In [82]:
```
d_s.query('int > 4') # extract quanitative values
```

Out[82]:

	int	str	U_str	float
4	5	e	E	1.5

In [83]:
```
d_s[d_s['int']>4]
```

Out[83]:

	int	str	U_str	float
4	5	e	E	1.5

In [84]:
```
d_s[['int','str','float']].query('int<6 & float>1.1')
```

Out[84]:

	int	str	float
1	2	b	1.2
2	3	c	1.3
3	4	d	1.4
4	5	e	1.5

In [85]:
```
d_s[['int','str','float']].query('int<6 and float>1.1')
```

Out[85]:

	int	str	float
1	2	b	1.2
2	3	c	1.3
3	4	d	1.4
4	5	e	1.5

In [86]:
```
d_s[['int','str','float']][(d_s['int']<6) & (d_s['float']>1.1)]
```

Out[86]:

	int	str	float
1	2	b	1.2
2	3	c	1.3
3	4	d	1.4
4	5	e	1.5

In [87]:
```
#d_s[['int','str','float']][(d_s['int']<6) and (d_s['float']>1.1)]
```

In [88]:
```
d_s.query('str=="c"')
```

Out[88]:

	int	str	U_str	float
2	3	c	C	1.3

In [89]:
```
d_s.query('int==3')
```

Out[89]:

	int	str	U_str	float
2	3	c	C	1.3

In [90]:
```
d_s
```

Out[90]:

	int	str	U_str	float
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

In [91]:
```
d_s[['int','str']][d_s['float']<1.9]
```

Out[91]:

	int	str
0	1	a
1	2	b
2	3	c
3	4	d
4	5	e

In [ ]:
```

```

In [92]:
```
d_l[['name','marks']][d_l['marks']>98]
```

Out[92]:

	name	marks
r1	om	100
r3	sharma	99

In [93]:
```
d_l[['name','marks']].query('marks>98')
```

Out[93]:

	name	marks
r1	om	100
r3	sharma	99

In [94]:
```
d_l[['name','marks']][d_l['marks']>97].sum()
```

Out[94]:
```
name      omkantsharma
marks              297
dtype: object
```

In [95]:
```
d_l[['name','marks']].query('marks>90').sum()
```

Out[95]:
```
name     omkantsharma
marks             297
dtype: object
```

In [96]:
```
d_l[['marks','name']][d_l['marks']==100]
```

Out[96]:

	marks	name
r1	100	om

In [97]:
```
d_l[['marks','name']].query('marks==100')
```

Out[97]:

	marks	name
r1	100	om

In [98]:
```
l=[33,32,31]
l_l=pd.Series(l,index=('r1','r2','r3'))
d_l['ages']=l_l
d_l
```

Out[98]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [99]:
```
d_l.query('marks>=100 and ages<=33')
```

Out[99]:

	name	marks	place	ages
r1	om	100	Indore	33

In [100]:
```
d_l.query('marks>=100 or ages<=33')
```

Out[100]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [101]:
```
d_l.query('(marks>=100 and ages>=33) and name=="om"')
```

Out[101]:

	name	marks	place	ages
r1	om	100	Indore	33

In [102]:
```
d_l.query('marks>=100 or ages<=33 or name=="om"')
```

Out[102]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [103]:
```
d_l[(d_l['marks']>=100) | (d_l['ages']<=33) | (d_l['name']=='om')]
```

Out[103]:

	name	marks	place	ages
r1	om	100	Indore	33
r2	kant	98	Ujjain	32
r3	sharma	99	Bhopal	31

In [104]:
```
df.query('PetalLengthCm>1 and PetalWidthCm>0.3 and Species=="Iris-setosa"')
```

Out[104]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5	6	5.4	3.9	1.7	0.4	Iris-setosa
15	16	5.7	4.4	1.5	0.4	Iris-setosa
16	17	5.4	3.9	1.3	0.4	Iris-setosa
21	22	5.1	3.7	1.5	0.4	Iris-setosa
23	24	5.1	3.3	1.7	0.5	Iris-setosa
26	27	5.0	3.4	1.6	0.4	Iris-setosa
31	32	5.4	3.4	1.5	0.4	Iris-setosa
43	44	5.0	3.5	1.6	0.6	Iris-setosa
44	45	5.1	3.8	1.9	0.4	Iris-setosa

In [105]:
```
df[(df['PetalLengthCm']>1) & (df['PetalWidthCm']>0.3) & (df['Species']=='Iris-setosa')]
```

Out[105]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
5	6	5.4	3.9	1.7	0.4	Iris-setosa
15	16	5.7	4.4	1.5	0.4	Iris-setosa
16	17	5.4	3.9	1.3	0.4	Iris-setosa
21	22	5.1	3.7	1.5	0.4	Iris-setosa
23	24	5.1	3.3	1.7	0.5	Iris-setosa
26	27	5.0	3.4	1.6	0.4	Iris-setosa
31	32	5.4	3.4	1.5	0.4	Iris-setosa
43	44	5.0	3.5	1.6	0.6	Iris-setosa
44	45	5.1	3.8	1.9	0.4	Iris-setosa

In [106]:
```
print(df['Species'].value_counts())
df_cq=df[df['Species']!='Iris-setosa']
print(df_cq['Species'].value_counts())
```

```
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
Iris-versicolor    50
Iris-virginica     50
Name: Species, dtype: int64
```

v.groupy()

In [107]:
```
df.columns
```

Out[107]:
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species'],
      dtype='object')
```

In [108]:
```
df
```

Out[108]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...	...
145	146	6.7	3.0	5.2	2.3	Iris-virginica
146	147	6.3	2.5	5.0	1.9	Iris-virginica
147	148	6.5	3.0	5.2	2.0	Iris-virginica
148	149	6.2	3.4	5.4	2.3	Iris-virginica
149	150	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 6 columns

```
df4=df.drop(['e',20,'201'],axis=1)
df4.head(2)
```

In [109]:
```
df4=df
```

In [110]:
```
df4.columns=['i','sl','sw','pl','pw','s']
```

In [111]:
```
df4.groupby(['s'])['sl'].sum()
```

Out[111]:
```
s
Iris-setosa        250.3
Iris-versicolor    296.8
Iris-virginica     329.4
Name: sl, dtype: float64
```

In [112]:
```
df4['sl'].groupby(df4['s']).sum()
```

Out[112]:
```
s
Iris-setosa        250.3
Iris-versicolor    296.8
Iris-virginica     329.4
Name: sl, dtype: float64
```

In [113]:
```
se=df4['sl'].groupby(df4['s']).sum()
```

In [114]:
```
print(se.index.tolist())
print(se.values.tolist())
```

```
['Iris-setosa', 'Iris-versicolor', 'Iris-virginica']
[250.3, 296.8, 329.4]
```

In [ ]:
```

```

Data Cleaning - Rename, Drop Columns, dropna, Fillna.ipynb

- df.drop('R_no./name') by defalt is axis 0
- df.drop('R_no./name',axis=0) - drop rows
- df.drop('Col_no./name'axis=1) - drop column copy
- df.drop(['col_name'],axis=1,inplace=True)
- df=df.drop(['B',"C"],axis=1)
- df.columns=['one','Two','Three','four','five','six','seven','eight']
- df.index=['a','b','c','d','e']

In [115]:
```
d_s.drop([1,2,3]) # drop row by indexing
```

Out[115]:

	int	str	U_str	float
0	1	a	A	1.1
4	5	e	E	1.5

In [116]:
```
d_s.drop(['str','float','int'],axis=1)
```

Out[116]:

	U_str
0	A
1	B
2	C
3	D
4	E

In [117]:
```
df.head(2)
```

Out[117]:

	i	sl	sw	pl	pw	s
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa

In [118]:
```
df=pd.read_csv('C:/Users/omkan/Desktop/Iris.csv')
```

In [119]:
```
df.drop(['SepalLengthCm','PetalLengthCm','PetalWidthCm'],axis=1).head()
```

Out[119]:

	Id	SepalWidthCm	Species
0	1	3.5	Iris-setosa
1	2	3.0	Iris-setosa
2	3	3.2	Iris-setosa
3	4	3.1	Iris-setosa
4	5	3.6	Iris-setosa

In [120]:
```
df.drop([1,3,5]).head(5)
```

Out[120]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa
6	7	4.6	3.4	1.4	0.3	Iris-setosa
7	8	5.0	3.4	1.5	0.2	Iris-setosa

In [121]:
```
# rename columns with list
d_s.columns=list('isuf')
d_s
```

Out[121]:

	i	s	u	f
0	1	a	A	1.1
1	2	b	B	1.2
2	3	c	C	1.3
3	4	d	D	1.4
4	5	e	E	1.5

In [122]:
```
d_s['f']=333  # add item in dataset as whole column
d_s.loc[0,'i']=123  # v.loc[row,'col']=value
d_s.loc[0,'s']=311
d_s
df['e']=100
df[20]=20
df.head()
df.loc[0,'201']=33
df.head()
```

Out[122]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	e	20	201
0	1	5.1	3.5	1.4	0.2	Iris-setosa	100	20	33.0
1	2	4.9	3.0	1.4	0.2	Iris-setosa	100	20	NaN
2	3	4.7	3.2	1.3	0.2	Iris-setosa	100	20	NaN
3	4	4.6	3.1	1.5	0.2	Iris-setosa	100	20	NaN
4	5	5.0	3.6	1.4	0.2	Iris-setosa	100	20	NaN

In [123]:
```
df.head(2)
```

Out[123]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species	e	20	201
0	1	5.1	3.5	1.4	0.2	Iris-setosa	100	20	33.0
1	2	4.9	3.0	1.4	0.2	Iris-setosa	100	20	NaN

In [ ]:
```
```

In [124]:
```
df.columns
```

Out[124]:
```
Index([          'Id', 'SepalLengthCm',  'SepalWidthCm', 'PetalLengthCm',
       'PetalWidthCm',       'Species',              'e',              20,
                  '201'],
      dtype='object')
```

In [125]:
```
df1=df.drop(['e',20,'201'],axis=1)
df1.head(2)
```

Out[125]:

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa

Rename columns

In [126]:
```
# rename all columns
df1.columns
```

Out[126]:
```
Index(['Id', 'SepalLengthCm', 'SepalWidthCm', 'PetalLengthCm', 'PetalWidthCm',
       'Species'],
      dtype='object')
```

In [127]:
```
# rename columns
df1.columns = ['id','sepallength','sepalwidth','petallength','petalwidth','species']
df1.head(1)
```

Out[127]:

	id	sepallength	sepalwidth	petallength	petalwidth	species
0	1	5.1	3.5	1.4	0.2	Iris-setosa

In [128]:
```
# rename specific columns
df1.rename(columns={'sepallength':'sl','sepalwidth':'sw','petallength':'pl','petalwidth':'pw'},inplace=True)
df1.head(2)
```

Out[128]:

	id	sl	sw	pl	pw	species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa

In [129]:
```
# rearrange columns
# method = 1
col=['sw','pl','pw','id','sl','species']
df1=pd.DataFrame(df1,columns=col)
df1.head(1)
```

Out[129]:

	sw	pl	pw	id	sl	species
0	3.5	1.4	0.2	1	5.1	Iris-setosa

In [130]:
```
# method -2
df1=df1[['id','sw','sl','pl','pw','species']]
df1.head(1)
```

Out[130]:

	id	sw	sl	pl	pw	species
0	1	3.5	5.1	1.4	0.2	Iris-setosa

drop the duplicate - df= drop_duplicate row

In [131]:
```
d_s[10]=20
d_s[10]=22
print(d_s.head(2))
```

```
     i    s  u    f  10
0  123  311  A  333  22
1    2    b  B  333  22
```

In [132]:
```
d_d=d_s.drop(['i','s','u','f'],axis=1)
d_d.head(2)
```

Out[132]:

	10
0	22
1	22

In [133]:
```
d_d=d_d.drop_duplicates() # row wise
d_d
```

Out[133]:

	10
0	22

In [134]:
```
d_s.drop_duplicates([10])  # drop value column_name wise
```

Out[134]:

	i	s	u	f	10
0	123	311	A	333	22

In [135]:
```
d_s.columns
```

Out[135]:
```
Index(['i', 's', 'u', 'f', 10], dtype='object')
```

In [136]:
```
d_s[5,10]=np.nan
print(d_s)
d_s.drop_duplicates([(5,10),10])
```

```
     i    s  u    f  10  (5, 10)
0  123  311  A  333  22       NaN
1    2    b  B  333  22       NaN
2    3    c  C  333  22       NaN
3    4    d  D  333  22       NaN
4    5    e  E  333  22       NaN
```

Out[136]:

	i	s	u	f	10	(5, 10)
0	123	311	A	333	22	NaN

In [137]:
```
# delete -
import sys
del d_s[(5,10)]
```

In [138]:
```
d_s.head(2)
```

Out[138]:

	i	s	u	f	10
0	123	311	A	333	22
1	2	b	B	333	22

v.dropna()

In [139]:
```
d_s[100]=np.nan
d_s.iloc[1,1]=np.nan
d_s.loc[3,'f']=np.nan
d_s[3,'u']=np.nan
d_s
```

Out[139]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN
1	2	NaN	B	333.0	22	NaN	NaN
2	3	c	C	333.0	22	NaN	NaN
3	4	d	D	NaN	22	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN

In [140]:
```
d_s.dropna()
```

Out[140]:

i	s	u	f	10	100	(3, u)

In [141]:
```
d_s.dropna(axis=0, how='any')  # row wise
```

Out[141]:

i	s	u	f	10	100	(3, u)

In [142]:
```
d_1=d_s.dropna(axis=1,how='all')  # drop column with all null
d_1
```

Out[142]:

	i	s	u	f	10
0	123	311	A	333.0	22
1	2	NaN	B	333.0	22
2	3	c	C	333.0	22
3	4	d	D	NaN	22
4	5	e	E	333.0	22

In [143]:
```
d_1.dropna(axis=0,how='any')      # drop row with any null
```

Out[143]:

	i	s	u	f	10
0	123	311	A	333.0	22
2	3	c	C	333.0	22
4	5	e	E	333.0	22

v.fillna()

In [144]:
```
d_s.head()
```

Out[144]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN
1	2	NaN	B	333.0	22	NaN	NaN
2	3	c	C	333.0	22	NaN	NaN
3	4	d	D	NaN	22	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN

In [145]:
```
d_s.columns
```

Out[145]:
```
Index(['i', 's', 'u', 'f', 10, 100, (3, 'u')], dtype='object')
```

In [146]:
```
d_s.rename( columns={(6,10):'e',(2,'u'):'g'},inplace=True)
d_s.head(1)
```

Out[146]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN

In [147]:
```
#d_s.columns=list('123456789')
#d_s
```

In [148]:
```
d_s.fillna(0).head(2)
```

Out[148]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	0.0	0.0
1	2	0	B	333.0	22	0.0	0.0

In [149]:
```
d_s.fillna(method='ffill') # forward fill
```

Out[149]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN
1	2	311	B	333.0	22	NaN	NaN
2	3	c	C	333.0	22	NaN	NaN
3	4	d	D	333.0	22	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN

In [150]:
```
d_s.fillna(method='bfill') # backword fill
```

Out[150]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN
1	2	c	B	333.0	22	NaN	NaN
2	3	c	C	333.0	22	NaN	NaN
3	4	d	D	333.0	22	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN

In [151]:
```
d_s
```

Out[151]:

	i	s	u	f	10	100	(3, u)
0	123	311	A	333.0	22	NaN	NaN
1	2	NaN	B	333.0	22	NaN	NaN
2	3	c	C	333.0	22	NaN	NaN
3	4	d	D	NaN	22	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN

fill mean very importent

In [152]:
```
d_s['int']=[11,12,13,np.nan,15]
d_s
```

Out[152]:

	i	s	u	f	10	100	(3, u)	int
0	123	311	A	333.0	22	NaN	NaN	11.0
1	2	NaN	B	333.0	22	NaN	NaN	12.0
2	3	c	C	333.0	22	NaN	NaN	13.0
3	4	d	D	NaN	22	NaN	NaN	NaN
4	5	e	E	333.0	22	NaN	NaN	15.0

In [153]:
```
m_1=d_s['int'].mean()
m_1
```

Out[153]: 12.75

In [154]:
```
#fill_values = {'A': df['A'].mean(), 'B': 34, 'C': 2, 'D': 3}
fill={'int':m_1,100:2,'f':3,(3,'u'):2}
d_f=d_s.fillna(value=fill)
d_f
```

Out[154]:

	i	s	u	f	10	100	(3, u)	int
0	123	311	A	333.0	22	2.0	2.0	11.00
1	2	NaN	B	333.0	22	2.0	2.0	12.00
2	3	c	C	333.0	22	2.0	2.0	13.00
3	4	d	D	3.0	22	2.0	2.0	12.75
4	5	e	E	333.0	22	2.0	2.0	15.00

In [155]:
```
d_f.columns
```

Out[155]: Index(['i', 's', 'u', 'f', 10, 100, (3, 'u'), 'int'], dtype='object')

In [ ]:
```
```

In [156]:
```
# v.apply(np.sum,axis=0)
# v.apply(np.sum,axis=1)
# v.apply(np.mean,axis=1)
```

Merge, Append, Concatinate

In [157]:
```
d1=pd.DataFrame({'a':[1,2,4,6],
                 'B':[12,13,14,15],
                 'C':[111,222,333,444]},index=('R1','R2','R3','R4'))
d2=pd.DataFrame({'a':[1,4,6,7],
                 'b':[1.2,1.3,1.4,1.5],
                 'c':[100,200,300,400]},index=('r1','r2','r3','r4'))
```

In [158]:
```
d1
```

Out[158]:

	a	B	C
R1	1	12	111
R2	2	13	222
R3	4	14	333
R4	6	15	444

In [159]:
```
d2
```

Out[159]:

	a	b	c
r1	1	1.2	100
r2	4	1.3	200
r3	6	1.4	300
r4	7	1.5	400

Merge

- inner : Return only the rows in which the left table has matching keys in the right table
- v=pd.merge(v1,v2,on='Based col_name', how=inner)

In [160]:
```
d=pd.merge(d1,d2,on='a',how='inner')
d
```

Out[160]:

	a	B	C	b	c
0	1	12	111	1.2	100
1	4	14	333	1.3	200
2	6	15	444	1.4	300

In [161]:
```
d=pd.merge(d1,d2,on='a',how='outer')
d
```

Out[161]:

	a	B	C	b	c
0	1	12.0	111.0	1.2	100.0
1	2	13.0	222.0	NaN	NaN
2	4	14.0	333.0	1.3	200.0
3	6	15.0	444.0	1.4	300.0
4	7	NaN	NaN	1.5	400.0

In [162]:
```
d=pd.merge(d1,d2,on='a',how='left')
d
```

Out[162]:

	a	B	C	b	c
0	1	12	111	1.2	100.0
1	2	13	222	NaN	NaN
2	4	14	333	1.3	200.0
3	6	15	444	1.4	300.0

In [163]:
```
d=pd.merge(d1,d2,on='a',how='right')
d
```

Out[163]:

	a	B	C	b	c
0	1	12.0	111.0	1.2	100
1	4	14.0	333.0	1.3	200
2	6	15.0	444.0	1.4	300
3	7	NaN	NaN	1.5	400

append

In [164]:
```
d=d1.append(d2)
d
```

Out[164]:

	a	B	C	b	c
R1	1	12.0	111.0	NaN	NaN
R2	2	13.0	222.0	NaN	NaN
R3	4	14.0	333.0	NaN	NaN
R4	6	15.0	444.0	NaN	NaN
r1	1	NaN	NaN	1.2	100.0
r2	4	NaN	NaN	1.3	200.0
r3	6	NaN	NaN	1.4	300.0
r4	7	NaN	NaN	1.5	400.0

In [165]:
```
d=d2.append(d1)
d
```

Out[165]:

	a	b	c	B	C
r1	1	1.2	100.0	NaN	NaN
r2	4	1.3	200.0	NaN	NaN
r3	6	1.4	300.0	NaN	NaN
r4	7	1.5	400.0	NaN	NaN
R1	1	NaN	NaN	12.0	111.0
R2	2	NaN	NaN	13.0	222.0
R3	4	NaN	NaN	14.0	333.0
R4	6	NaN	NaN	15.0	444.0

In [166]:
```
d.columns
d.index=list(range(0,8))
d
```

Out[166]:

	a	b	c	B	C
0	1	1.2	100.0	NaN	NaN
1	4	1.3	200.0	NaN	NaN
2	6	1.4	300.0	NaN	NaN
3	7	1.5	400.0	NaN	NaN
4	1	NaN	NaN	12.0	111.0
5	2	NaN	NaN	13.0	222.0
6	4	NaN	NaN	14.0	333.0
7	6	NaN	NaN	15.0	444.0

In [167]:
```
# concat
pd.concat([d1,d2])
```

Out[167]:

	a	B	C	b	c
R1	1	12.0	111.0	NaN	NaN
R2	2	13.0	222.0	NaN	NaN
R3	4	14.0	333.0	NaN	NaN
R4	6	15.0	444.0	NaN	NaN
r1	1	NaN	NaN	1.2	100.0
r2	4	NaN	NaN	1.3	200.0
r3	6	NaN	NaN	1.4	300.0
r4	7	NaN	NaN	1.5	400.0

In [168]:
```
pd.concat([d2,d1],axis=1)
```

Out[168]:

	a	b	c	a	B	C
r1	1.0	1.2	100.0	NaN	NaN	NaN
r2	4.0	1.3	200.0	NaN	NaN	NaN
r3	6.0	1.4	300.0	NaN	NaN	NaN
r4	7.0	1.5	400.0	NaN	NaN	NaN
R1	NaN	NaN	NaN	1.0	12.0	111.0
R2	NaN	NaN	NaN	2.0	13.0	222.0
R3	NaN	NaN	NaN	4.0	14.0	333.0
R4	NaN	NaN	NaN	6.0	15.0	444.0

Transpose

In [169]:
```
df.T
```

Out[169]:

	0	1	2	3	4	5	6	7	8	9	...	140	141	142	143	
Id	1	2	3	4	5	6	7	8	9	10	...	141	142	143	144	
SepalLengthCm	5.1	4.9	4.7	4.6	5.0	5.4	4.6	5.0	4.4	4.9	...	6.7	6.9	5.8	6.8	
SepalWidthCm	3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1	...	3.1	3.1	2.7	3.2	
PetalLengthCm	1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5	...	5.6	5.1	5.1	5.9	
PetalWidthCm	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2	0.2	0.1	...	2.4	2.3	1.9	2.3	
Species	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	Iris-setosa	...	Iris-virginica	Iris-virginica	Iris-virginica	Iris-virginica	v
e	100	100	100	100	100	100	100	100	100	100	...	100	100	100	100	
20	20	20	20	20	20	20	20	20	20	20	...	20	20	20	20	
201	33.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	

9 rows × 150 columns

In [ ]:
```
```

In [170]:
```
#rediction
df
d_l=df
d_l
d_l.mean(),d_l.median(),d_l.std(),d_l.var() # ,d_l.mode()
```

```
C:\Users\omkan\AppData\Local\Temp/ipykernel_16548/1306969315.py:5: FutureWarning: Dropping of nuisance colu
mns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise T
ypeError.  Select only valid columns before calling the reduction.
  d_l.mean(),d_l.median(),d_l.std(),d_l.var() # ,d_l.mode()
```

Out[170]:
```
(Id                75.500000
 SepalLengthCm      5.843333
 SepalWidthCm       3.054000
 PetalLengthCm      3.758667
 PetalWidthCm       1.198667
 e                100.000000
 20                20.000000
 201               33.000000
 dtype: float64,
 Id                75.50
 SepalLengthCm      5.80
 SepalWidthCm       3.00
 PetalLengthCm      4.35
 PetalWidthCm       1.30
 e                100.00
 20                20.00
 201               33.00
 dtype: float64,
 Id                43.445368
 SepalLengthCm      0.828066
 SepalWidthCm       0.433594
 PetalLengthCm      1.764420
 PetalWidthCm       0.763161
 e                  0.000000
 20                 0.000000
 201                     NaN
 dtype: float64,
 Id                1887.500000
 SepalLengthCm        0.685694
 SepalWidthCm         0.188004
 PetalLengthCm        3.113179
 PetalWidthCm         0.582414
 e                    0.000000
 20                   0.000000
 201                       NaN
 dtype: float64)
```

DataFrame statistical functions -

- mean(),median(),mode(),std(),var()

In [171]:
```
df.mean()
```

```
C:\Users\omkan\AppData\Local\Temp/ipykernel_16548/3698961737.py:1: FutureWarning: Dropping of nuisance colu
mns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise T
ypeError.  Select only valid columns before calling the reduction.
  df.mean()
```

Out[171]:
```
Id               75.500000
SepalLengthCm     5.843333
SepalWidthCm      3.054000
PetalLengthCm     3.758667
PetalWidthCm      1.198667
e               100.000000
20               20.000000
201              33.000000
dtype: float64
```

In [172]:
```
df.median()
```

```
C:\Users\omkan\AppData\Local\Temp/ipykernel_16548/530051474.py:1: FutureWarning: Dropping of nuisance colum
ns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise Ty
peError.  Select only valid columns before calling the reduction.
  df.median()
```

Out[172]:
```
Id               75.50
SepalLengthCm     5.80
SepalWidthCm      3.00
PetalLengthCm     4.35
PetalWidthCm      1.30
e               100.00
20               20.00
201              33.00
dtype: float64
```

In [173]:
```
#df.mode()
df.std()
```

```
C:\Users\omkan\AppData\Local\Temp/ipykernel_16548/300518187.py:2: FutureWarning: Dropping of nuisance colum
ns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise Ty
peError.  Select only valid columns before calling the reduction.
  df.std()
```

Out[173]:
```
Id               43.445368
SepalLengthCm     0.828066
SepalWidthCm      0.433594
PetalLengthCm     1.764420
PetalWidthCm      0.763161
e                 0.000000
20                0.000000
201                    NaN
dtype: float64
```

In [174]:
```
df.var()  #var() function that calculates the variance along all columns
```

```
C:\Users\omkan\AppData\Local\Temp/ipykernel_16548/2027085253.py:1: FutureWarning: Dropping of nuisance colu
mns in DataFrame reductions (with 'numeric_only=None') is deprecated; in a future version this will raise T
ypeError.  Select only valid columns before calling the reduction.
  df.var()  #var() function that calculates the variance along all columns
```

Out[174]:
```
Id               1887.500000
SepalLengthCm       0.685694
SepalWidthCm        0.188004
PetalLengthCm       3.113179
PetalWidthCm        0.582414
e                   0.000000
20                  0.000000
201                      NaN
dtype: float64
```

In [ ]:
```
```