Assignment No.1

Aim :

1. Introduction to Dataset
2. Python Libraries for Data Science
3. Description of Dataset
4. Panda Dataframe functions for load the dataset
5. Panda functions for Data Preprocessing
6. Panda functions for Data Formatting and Normalisation
7. Panda Functions for handling categorical variables

In [17]:
```
import pandas as pd
```

In [18]:
```
import seaborn as sns
```

In [19]:
```
import numpy as np
```

In [20]:
```
import matplotlib.pyplot as plt
```

In [21]:
```
data_set_name=sns.get_dataset_names()
```

In [22]:
```
print(data_set_name)
```

```
['anagrams', 'anscombe', 'attention', 'brain_networks', 'car_crashes', 'diamo
nds', 'dots', 'dowjones', 'exercise', 'flights', 'fmri', 'geyser', 'glue', 'h
ealthexp', 'iris', 'mpg', 'penguins', 'planets', 'seaice', 'taxis', 'tips',
'titanic', 'anagrams', 'anagrams', 'anscombe', 'anscombe', 'attention', 'atte
ntion', 'brain_networks', 'brain_networks', 'car_crashes', 'car_crashes', 'di
amonds', 'diamonds', 'dots', 'dots', 'dowjones', 'dowjones', 'exercise', 'exe
rcise', 'flights', 'flights', 'fmri', 'fmri', 'geyser', 'geyser', 'glue', 'gl
ue', 'healthexp', 'healthexp', 'iris', 'iris', 'mpg', 'mpg', 'penguins', 'pen
guins', 'planets', 'planets', 'seaice', 'seaice', 'taxis', 'taxis', 'tips',
'tips', 'titanic', 'titanic', 'anagrams', 'anscombe', 'attention', 'brain_net
works', 'car_crashes', 'diamonds', 'dots', 'dowjones', 'exercise', 'flights',
'fmri', 'geyser', 'glue', 'healthexp', 'iris', 'mpg', 'penguins', 'planets',
'seaice', 'taxis', 'tips', 'titanic']
```

In [23]:
```
dataset=sns.load_dataset("iris")
dataset
```

Out[23]:

	sepal_length	sepal_width	petal_length	petal_width	species
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 [40]:
```
dataset.head(6)
```

Out[40]:

	sepal_length	sepal_width	petal_length	petal_width	species
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
5	5.4	3.9	1.7	0.4	setosa

In [25]:
```
dataset.head(5)
```

Out[25]:

	sepal_length	sepal_width	petal_length	petal_width	species
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 [26]:
```
dataset.tail(5)
```

Out[26]:

	sepal_length	sepal_width	petal_length	petal_width	species
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

In [27]:
```
dataset.index
```

Out[27]:
```
RangeIndex(start=0, stop=150, step=1)
```

In [28]:
```
dataset.columns
```

Out[28]:
```
Index(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'],
      dtype='object')
```

In [33]:
```
dataset.shape
```

Out[33]:
```
(150, 5)
```

In [30]:
```
dataset.dtypes
```

Out[30]:
```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [31]:
```
dataset.columns.values
```

Out[31]:
```
array(['sepal_length', 'sepal_width', 'petal_length', 'petal_width',
       'species'], dtype=object)
```

In [34]:
```
dataset.describe(include='all')
```

Out[34]:

	sepal_length	sepal_width	petal_length	petal_width	species
count	150.000000	150.000000	150.000000	150.000000	150
unique	NaN	NaN	NaN	NaN	3
top	NaN	NaN	NaN	NaN	setosa
freq	NaN	NaN	NaN	NaN	50
mean	5.843333	3.057333	3.758000	1.199333	NaN
std	0.828066	0.435866	1.765298	0.762238	NaN
min	4.300000	2.000000	1.000000	0.100000	NaN
25%	5.100000	2.800000	1.600000	0.300000	NaN
50%	5.800000	3.000000	4.350000	1.300000	NaN
75%	6.400000	3.300000	5.100000	1.800000	NaN
max	7.900000	4.400000	6.900000	2.500000	NaN

In [35]:
```
dataset['sepal_width']
```

Out[35]:
```
0      3.5
1      3.0
2      3.2
3      3.1
4      3.6
      ...
145    3.0
146    2.5
147    3.0
148    3.4
149    3.0
Name: sepal_width, Length: 150, dtype: float64
```

In [37]:
```
dataset.sort_index(axis=1,ascending=0)
```

Out[37]:

	species	sepal_width	sepal_length	petal_width	petal_length
0	setosa	3.5	5.1	0.2	1.4
1	setosa	3.0	4.9	0.2	1.4
2	setosa	3.2	4.7	0.2	1.3
3	setosa	3.1	4.6	0.2	1.5
4	setosa	3.6	5.0	0.2	1.4
...	...	...	...	...	...
145	virginica	3.0	6.7	2.3	5.2
146	virginica	2.5	6.3	1.9	5.0
147	virginica	3.0	6.5	2.0	5.2
148	virginica	3.4	6.2	2.3	5.4
149	virginica	3.0	5.9	1.8	5.1

150 rows × 5 columns

In [38]:
```
dataset.sort_values(by="sepal_length")
```

Out[38]:

	sepal_length	sepal_width	petal_length	petal_width	species
13	4.3	3.0	1.1	0.1	setosa
42	4.4	3.2	1.3	0.2	setosa
38	4.4	3.0	1.3	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
41	4.5	2.3	1.3	0.3	setosa
...	...	...	...	...	...
122	7.7	2.8	6.7	2.0	virginica
118	7.7	2.6	6.9	2.3	virginica
117	7.7	3.8	6.7	2.2	virginica
135	7.7	3.0	6.1	2.3	virginica
131	7.9	3.8	6.4	2.0	virginica

150 rows × 5 columns

In [39]:
```
dataset.iloc[5]
```

Out[39]:
```
sepal_length       5.4
sepal_width        3.9
petal_length       1.7
petal_width        0.4
species         setosa
Name: 5, dtype: object
```

In [41]:
```
dataset[0:3]
```

Out[41]:

	sepal_length	sepal_width	petal_length	petal_width	species
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

In [44]:
```
dataset.loc[:,["sepal_length","sepal_width"]]
```

Out[44]:

	sepal_length	sepal_width
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6
5	5.4	3.9
6	4.6	3.4
...	...	...
145	6.7	3.0
146	6.3	2.5
147	6.5	3.0
148	6.2	3.4
149	5.9	3.0

148 rows × 2 columns

In [45]:
```
dataset.iloc[:4,:]
```

Out[45]:

	sepal_length	sepal_width	petal_length	petal_width	species
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

In [47]:
```
dataset.iloc[:,:2]
```

Out[47]:

	sepal_length	sepal_width
0	5.1	3.5
1	4.9	3.0
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6
...	...	...
145	6.7	3.0
146	6.3	2.5
147	6.5	3.0
148	6.2	3.4
149	5.9	3.0

150 rows × 2 columns

In [48]:
```
dataset.iloc[:5,:2]
```

Out[48]:

	sepal_length	sepal_width
0	5.1	3.5
1	4.9	3.0
2	4.7	3.2
3	4.6	3.1
4	5.0	3.6

In [50]:
```
dataset.iloc[3:5,0:3]
```

Out[50]:

	sepal_length	sepal_width	petal_length
3	4.6	3.1	1.5
4	5.0	3.6	1.4

In [52]:
```
dataset.iloc[[1,2,4],[0,2]]
```

Out[52]:

	sepal_length	petal_length
1	4.9	1.4
2	4.7	1.3
4	5.0	1.4

In [54]:
```
dataset.iloc[[1,9,10],[0,3]]
```

Out[54]:

	sepal_length	petal_width
1	4.9	0.2
9	4.9	0.1
10	5.4	0.2

In [55]:
```
dataset.iloc[1:3,:]
```

Out[55]:

	sepal_length	sepal_width	petal_length	petal_width	species
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa

In [56]:
```
dataset.iloc[:,1:3]
```

Out[56]:

	sepal_width	petal_length
0	3.5	1.4
1	3.0	1.4
2	3.2	1.3
3	3.1	1.5
4	3.6	1.4
...	...	...
145	3.0	5.2
146	2.5	5.0
147	3.0	5.2
148	3.4	5.4
149	3.0	5.1

150 rows × 2 columns

In [59]:
```
dataset.iloc[2,1]
```

Out[59]:
```
3.2
```

In [60]:
```
dataset["sepal_length"].iloc[5]
```

Out[60]:
```
5.4
```

In [62]:
```
c=dataset.columns[1:3]
dataset[c]
```

Out[62]:

	sepal_width	petal_length
0	3.5	1.4
1	3.0	1.4
2	3.2	1.3
3	3.1	1.5
4	3.6	1.4
...	...	...
145	3.0	5.2
146	2.5	5.0
147	3.0	5.2
148	3.4	5.4
149	3.0	5.1

150 rows × 2 columns

In [63]:
```
dataset[dataset.columns[2:4]].iloc[5:10]
```

Out[63]:

	petal_length	petal_width
5	1.7	0.4
6	1.4	0.3
7	1.5	0.2
8	1.4	0.2
9	1.5	0.1

NAME: PRITAM PARADE ROLLNO: 13257