Assignment No 1

In [2]:

```
# Importing all the required python libraries
import pandas as pd
import numpy as np
```

In [3]:

```
# Loading dataset into pandas data frame
df=pd.read_csv("iris.csv")
df
```

Out[3]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	3.5	1.4	0.2	setosa
1	4.9	NaN	1.4	0.2	setosa
2	4.7	3.2	NaN	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	NaN	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

In [4]:

```
# Data Preprocessing
# Descibe Function
df.describe()
```

Out[4]:

	sepal_length	sepal_width	petal_length	petal_width
count	146.000000	146.000000	146.000000	146.000000
mean	5.858219	3.045890	3.823288	1.224658
std	0.832508	0.432654	1.743878	0.756905
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.325000
50%	5.800000	3.000000	4.400000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

In [5]:

```
# Check dimension of data frame
df.shape
```

Out[5]:

```
(150, 5)
```

In [6]:

```
df.head()
```

Out[6]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	3.5	1.4	0.2	setosa
1	4.9	NaN	1.4	0.2	setosa
2	4.7	3.2	NaN	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	NaN	1.4	0.2	setosa

In [7]:

```
df.tail()
```

Out[7]:

	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 [8]:

```
df.head(2)
```

Out[8]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	3.5	1.4	0.2	setosa
1	4.9	NaN	1.4	0.2	setosa

In [9]:

```
df.tail(2)
```

Out[9]:

	sepal_length	sepal_width	petal_length	petal_width	species
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

In [10]:

```
# Types of variables
df.dtypes
```

Out[10]:

```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [11]:

```
# Check for missing values in data frame
df.isnull()
```

Out[11]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	True	False	False	False	False
1	False	True	False	False	False
2	False	False	True	False	False
3	False	False	False	False	False
4	False	True	False	False	False
...	...	...	...	...	...
145	False	False	False	False	False
146	False	False	False	False	False
147	False	False	False	False	False
148	False	False	False	False	False
149	False	False	False	False	False

150 rows × 5 columns

In [12]:

```
df.isnull().sum()
```

Out[12]:

```
sepal_length    4
sepal_width     4
petal_length    4
petal_width     4
species         0
dtype: int64
```

In [13]:

```
df.isnull().sum().sum()
```

Out[13]:

```
16
```

In [14]:

```
# Fill null value
df=df.fillna(value=0)
df
```

Out[14]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	0.0	3.5	1.4	0.2	setosa
1	4.9	0.0	1.4	0.2	setosa
2	4.7	3.2	0.0	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	0.0	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 [15]:

```
df=pd.read_csv("iris.csv")
df=df.fillna(method='pad')
df
```

Out[15]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	3.5	1.4	0.2	setosa
1	4.9	3.5	1.4	0.2	setosa
2	4.7	3.2	1.4	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.1	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 [16]:

```
df=pd.read_csv("iris.csv")
df=df.fillna(method='ffill')
df
```

Out[16]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	3.5	1.4	0.2	setosa
1	4.9	3.5	1.4	0.2	setosa
2	4.7	3.2	1.4	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.1	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 [17]:

```
df=pd.read_csv("iris.csv")
df=df.fillna(method='bfill')
df
```

Out[17]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	4.9	3.5	1.4	0.2	setosa
1	4.9	3.2	1.4	0.2	setosa
2	4.7	3.2	1.5	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.9	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 [18]:

```
# Fill null values with mean
df=pd.read_csv("iris.csv")
mean_value=df['sepal_length'].mean()
df['sepal_length'].fillna(value=mean_value,inplace=True)
df
```

Out[18]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.858219	3.5	1.4	0.2	setosa
1	4.900000	NaN	1.4	0.2	setosa
2	4.700000	3.2	NaN	0.2	setosa
3	4.600000	3.1	1.5	0.2	setosa
4	5.000000	NaN	1.4	0.2	setosa
...	...	...	...	...	...
145	6.700000	3.0	5.2	2.3	virginica
146	6.300000	2.5	5.0	1.9	virginica
147	6.500000	3.0	5.2	2.0	virginica
148	6.200000	3.4	5.4	2.3	virginica
149	5.900000	3.0	5.1	1.8	virginica

150 rows × 5 columns

In [19]:

```
# Fill null values with median
df=pd.read_csv("iris.csv")
median_value=df['sepal_length'].median()
df['sepal_length'].fillna(value=median_value,inplace=True)
df
```

Out[19]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.8	3.5	1.4	0.2	setosa
1	4.9	NaN	1.4	0.2	setosa
2	4.7	3.2	NaN	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	NaN	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 [20]:

```
# Fill null values with mode
df=pd.read_csv("iris.csv")
mode_value=df['sepal_length'].mode()
df['sepal_length'].fillna(value=mode_value,inplace=True)
df
```

Out[20]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.0	3.5	1.4	0.2	setosa
1	4.9	NaN	1.4	0.2	setosa
2	4.7	3.2	NaN	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	NaN	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 [21]:

```
# Data formatting and data normalization
df.dtypes
```

Out[21]:

```
sepal_length    float64
sepal_width     float64
petal_length    float64
petal_width  float64
species           object
dtype: object
```

In [22]:

```
# Changing data type of column
# Before changing data type all the null value must be handled properly else it will cre

mean_value=df['sepal_length'].mean()
df['sepal_length'].fillna(value=mean_value,inplace=True)
df['sepal_length']=df['sepal_length'].astype(int)

df.dtypes
```

Out[22]:

```
sepal_length      int32
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [23]:

```
mean_value=df['sepal_length'].mean()
df['sepal_length'].fillna(value=mean_value,inplace=True)

df['sepal_length']=df['sepal_length'].round(0).astype(int)
df.dtypes
```

Out[23]:

```
sepal_length      int32
sepal_width     float64
petal_length    float64
petal_width     float64
species          object
dtype: object
```

In [24]:

```
# Converting categorical variables into quantitative variables
df['species'].replace({'setosa':1,'versicolor':2,'virginica':3},inplace=True)
df
```

Out[24]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5	3.5	1.4	0.2	1
1	4	NaN	1.4	0.2	1
2	4	3.2	NaN	0.2	1
3	4	3.1	1.5	0.2	1
4	5	NaN	1.4	0.2	1
...	...	...	...	...	...
145	6	3.0	5.2	2.3	3
146	6	2.5	5.0	1.9	3
147	6	3.0	5.2	2.0	3
148	6	3.4	5.4	2.3	3
149	5	3.0	5.1	1.8	3

150 rows × 5 columns