Practical No. 02

Aim : To perform and analysis for Normal Distribution on given dataset

In [6]:

```
#Name : Taufiq Rafik Nagori
#Roll no. : 77 (BDA-B77)
#Section : B
#Subject : PE-II
```

In [8]:

```
import os
import pandas as pd
```

In [10]:

```
os.getcwd()
```

Out[10]:

```
'C:\\Users\\USER'
```

In [12]:

```
os.chdir("C:\\Users\\USER\\Desktop")
```

In [14]:

```
data = pd.read_csv("diabetes.csv")
```

In [16]:

```
data.head()
```

Out[16]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	6	148	72	35	0	33.6	0.627	50
1	1	85	66	29	0	26.6	0.351	31
2	8	183	64	0	0	23.3	0.672	32
3	1	89	66	23	94	28.1	0.167	21
4	0	137	40	35	168	43.1	2.288	33

In [18]:

```
data.tail()
```

Out[18]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Ag
763	10	101	76	48	180	32.9	0.171	6
764	2	122	70	27	0	36.8	0.340	2
765	5	121	72	23	112	26.2	0.245	3
766	1	126	60	0	0	30.1	0.349	4
767	1	93	70	31	0	30.4	0.315	2

In [20]:

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Pregnancies               768 non-null    int64  
 1   Glucose                   768 non-null    int64  
 2   BloodPressure             768 non-null    int64  
 3   SkinThickness             768 non-null    int64  
 4   Insulin                   768 non-null    int64  
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64  
 8   Outcome                   768 non-null    int64  
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

In [22]:

```
data.describe()
```

Out[22]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigr
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	

In [24]:

```
data.shape
```

Out[24]:

```
(768, 9)
```

In [26]:

```
data.size
```

Out[26]:

```
6912
```

In [28]:

```
data.ndim
```

Out[28]:

```
2
```

In [30]:

```
data.columns
```

Out[30]:

```
Index(['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin',
       'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome'],
      dtype='object')
```

Data pre-processing, data-cleaning, mising value treatment

In [33]:

```
data.isna()
```

Out[33]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	A
0	False	False	False	False	False	False	False	Fa
1	False	False	False	False	False	False	False	Fa
2	False	False	False	False	False	False	False	Fa
3	False	False	False	False	False	False	False	Fa
4	False	False	False	False	False	False	False	Fa
...	...	...	...	...	...	...	...	
763	False	False	False	False	False	False	False	Fa
764	False	False	False	False	False	False	False	Fa
765	False	False	False	False	False	False	False	Fa
766	False	False	False	False	False	False	False	Fa
767	False	False	False	False	False	False	False	Fa

768 rows × 9 columns

In [35]:

```
data.isna().any()
```

Out[35]:

```
Pregnancies                 False
Glucose                     False
BloodPressure               False
SkinThickness               False
Insulin                     False
BMI                         False
DiabetesPedigreeFunction    False
Age                         False
Outcome                     False
dtype: bool
```

In [37]:

```
data.isna().sum()
```

Out[37]:

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
```

```
Age                            0
Outcome                        0
dtype: int64
```

In [39]:

```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [41]:

```
sns.distplot(data,bins=20)
plt.show()
```

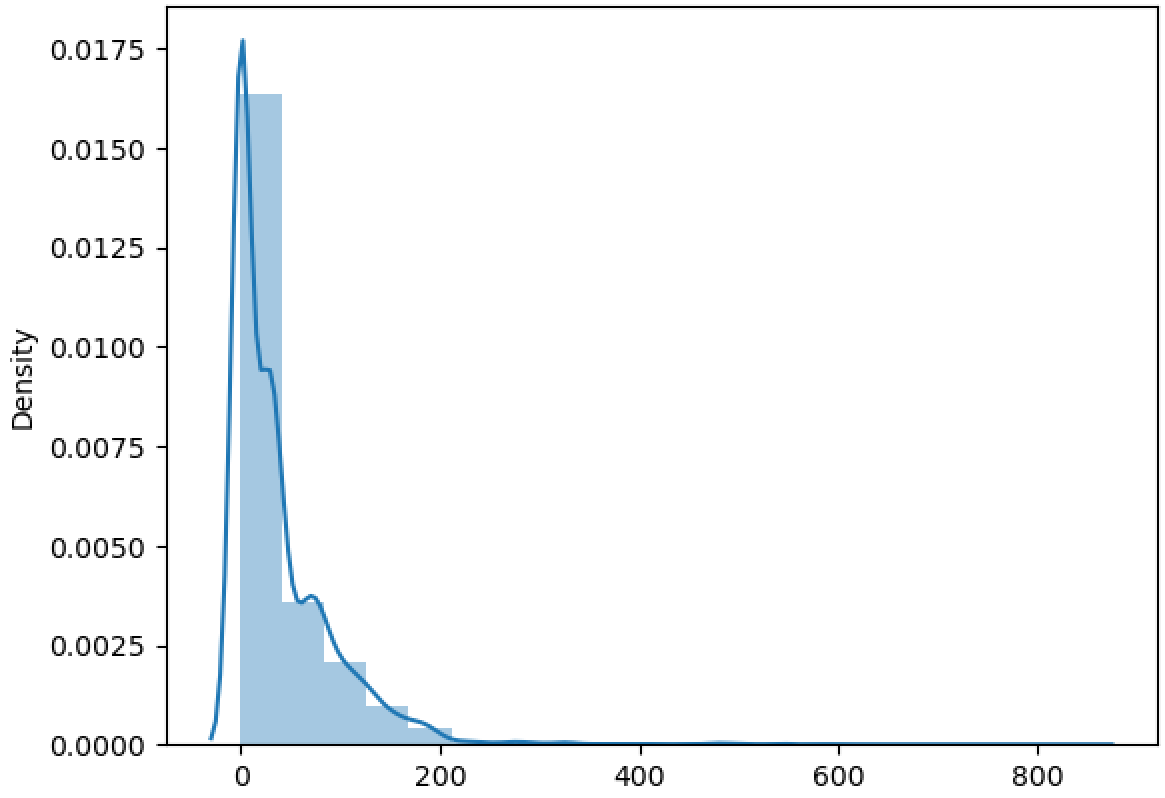
```
C:\Users\USER\AppData\Local\Temp\ipykernel_15968\1706651633.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data,bins=20)
```



In [43]:

```
sns.distplot(data['Glucose'],bins=20)
plt.show()
```

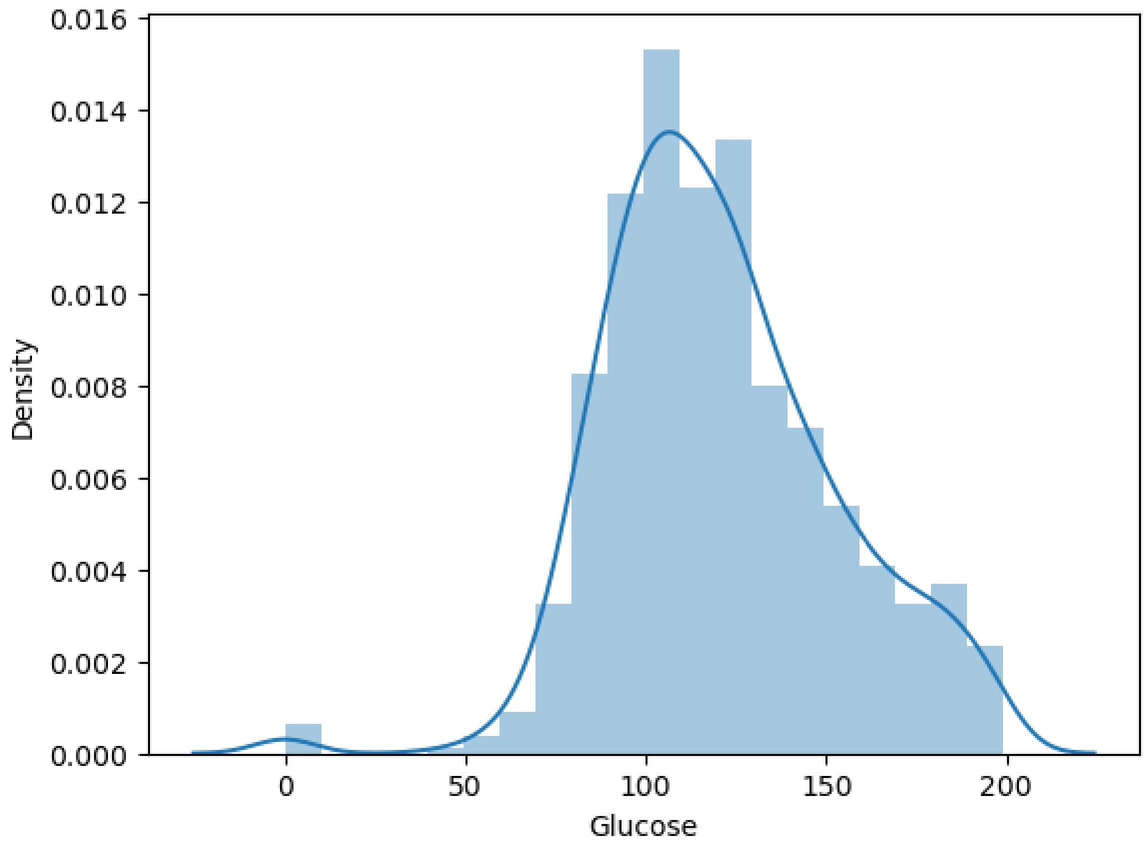
```
C:\Users\USER\AppData\Local\Temp\ipykernel_15968\1093375177.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).
```

```
For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data['Glucose'],bins=20)
```



In [45]:

```
sns.distplot(data['Age'], bins=20)
plt.show()
```

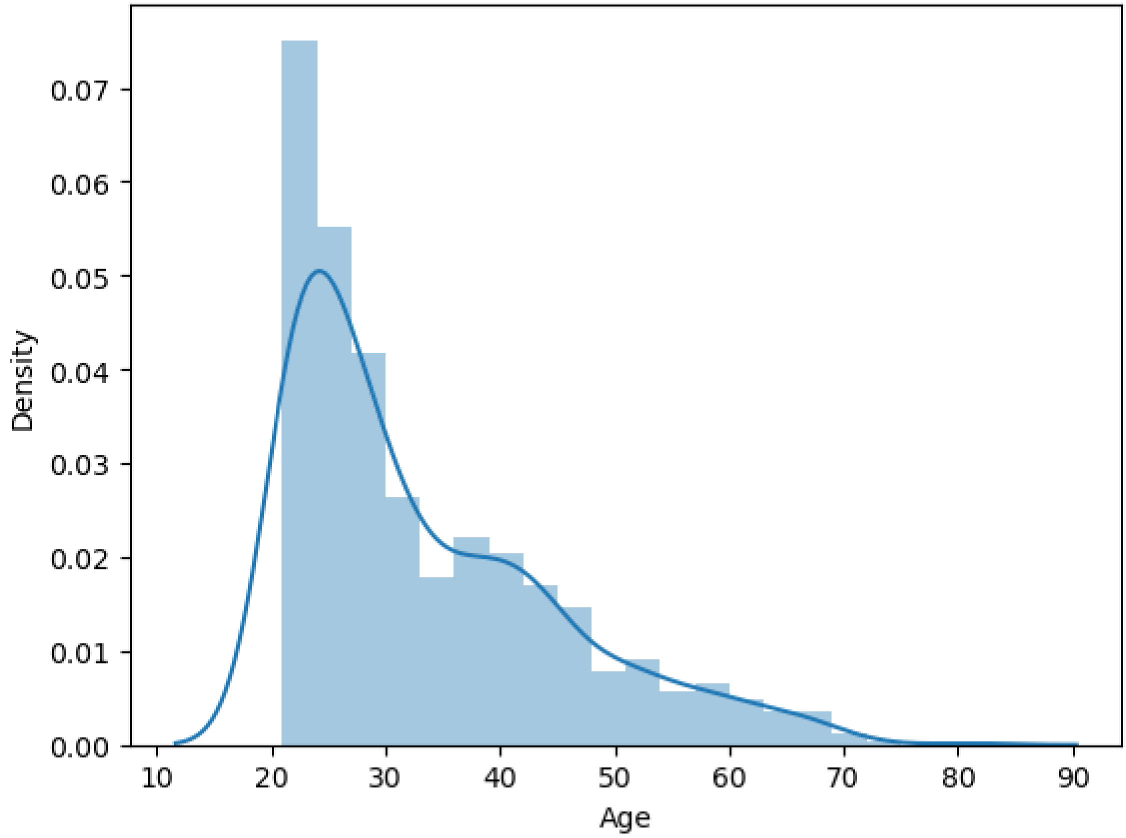
```
C:\Users\USER\AppData\Local\Temp\ipykernel_15968\713659080.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data['Age'], bins=20)
```



In [47]:

```
sns.distplot(data['BloodPressure'], bins = 20)
plt.show()
```

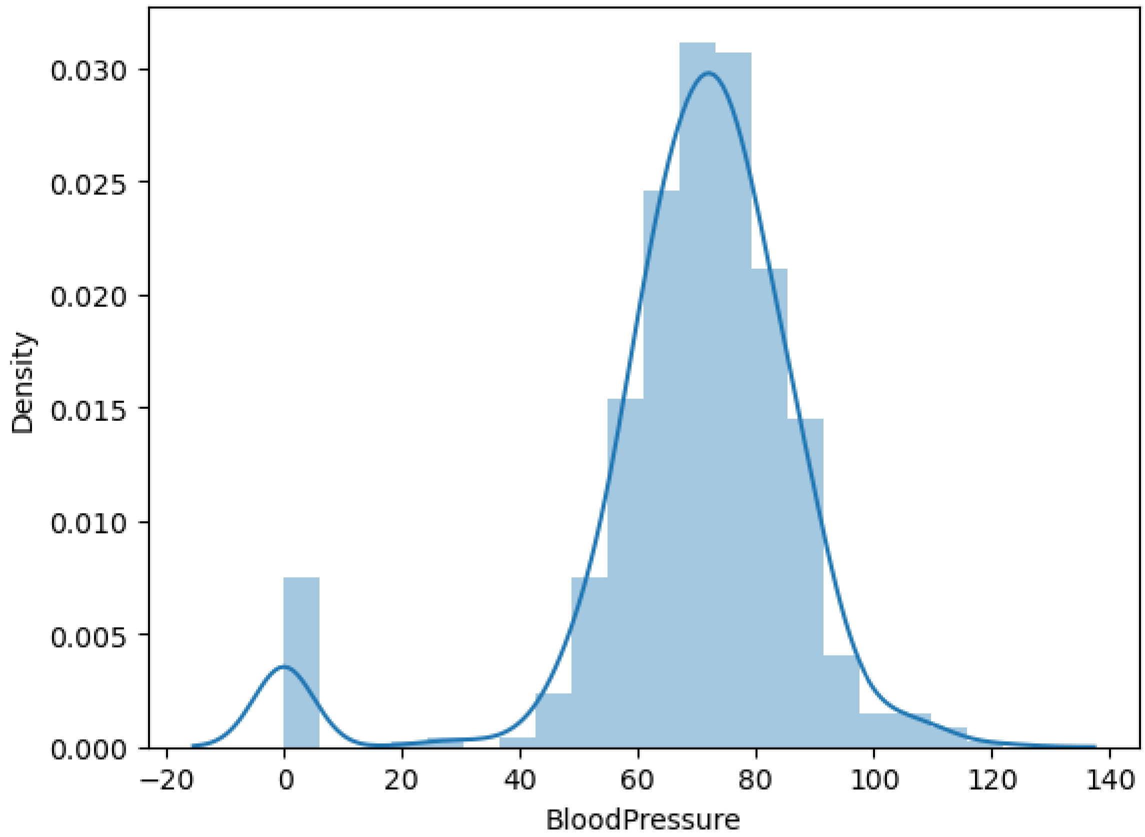
```
C:\Users\USER\AppData\Local\Temp\ipykernel_15968\1074119919.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data['BloodPressure'], bins = 20)
```



In [49]:

```
sns.distplot(data['SkinThickness'], bins = 20)
plt.show()
```

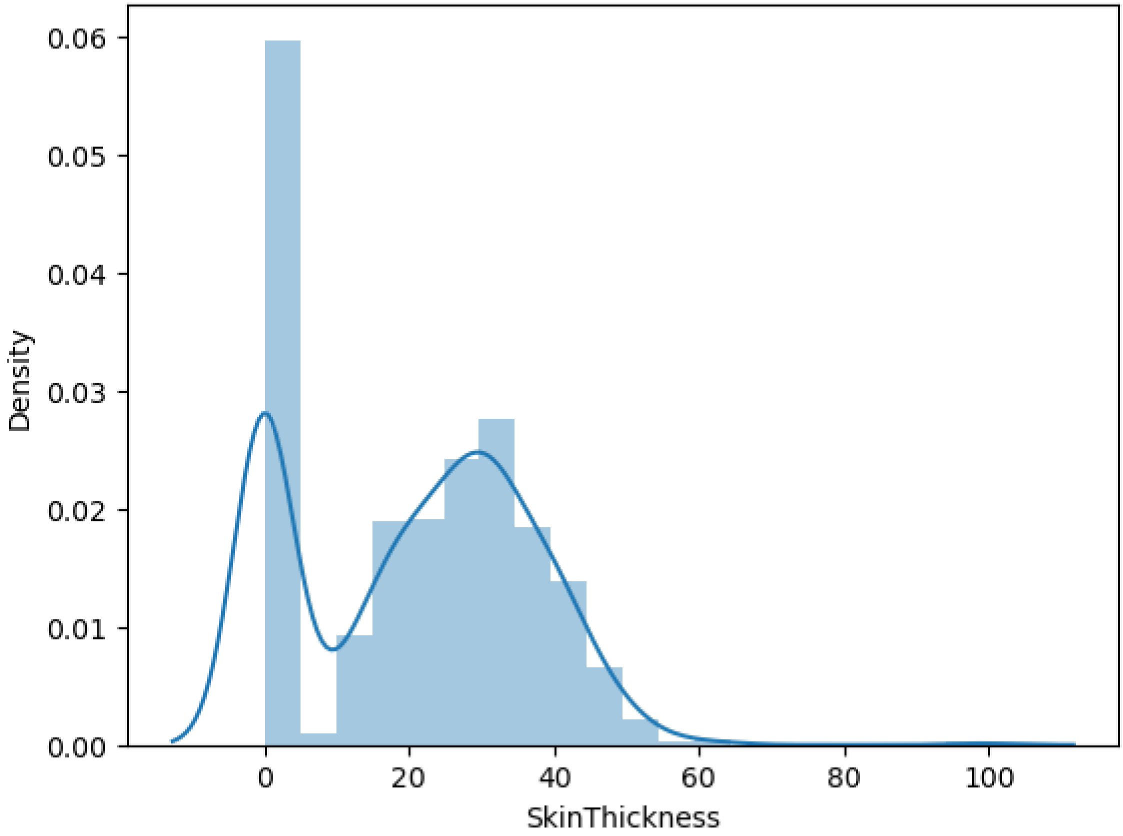
```
C:\Users\USER\AppData\Local\Temp\ipykernel_15968\3091487386.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(data['SkinThickness'], bins = 20)
```

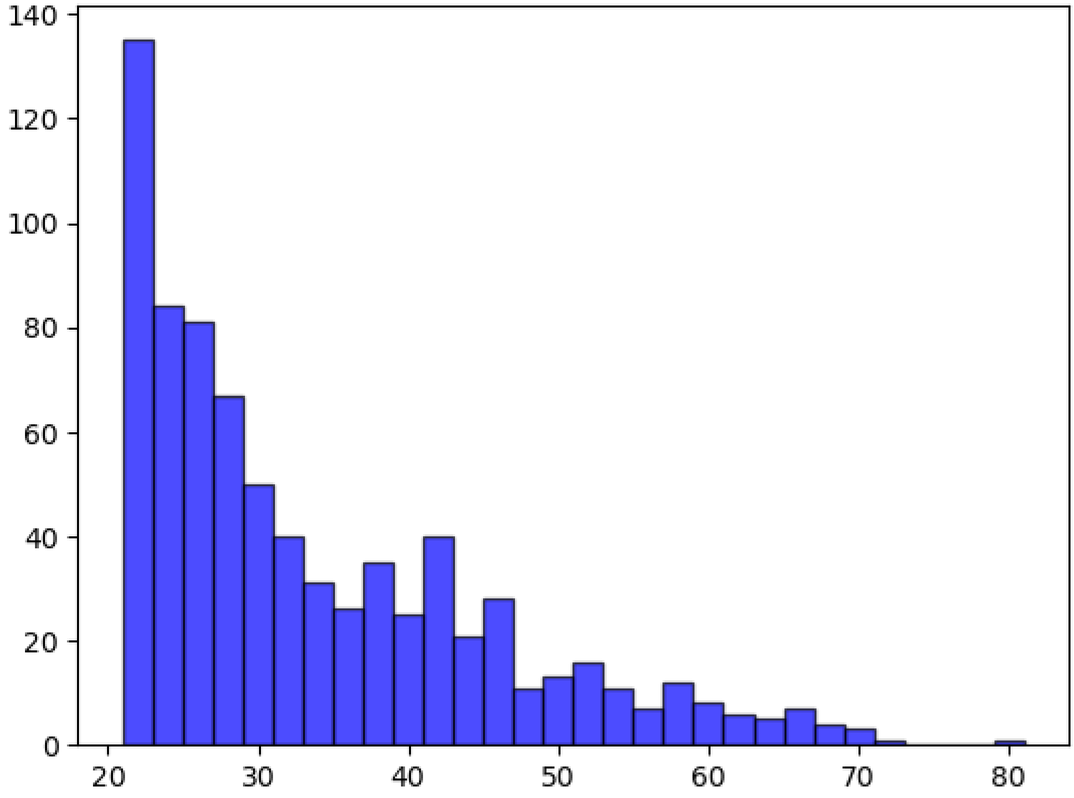


In [51]:

```
plt.hist(data['Age'], bins=30, color='blue', edgecolor='black', alpha=0.7)
```

Out[51]:

```
(array([135.,  84.,  81.,  67.,  50.,  40.,  31.,  26.,  35.,  25.,  40.,
         21.,  28.,  11.,  13.,  16.,  11.,   7.,  12.,   8.,   6.,   5.,
          7.,   4.,   3.,   1.,   0.,   0.,   0.,   1.]),
 array([21., 23., 25., 27., 29., 31., 33., 35., 37., 39., 41., 43., 45.,
        47., 49., 51., 53., 55., 57., 59., 61., 63., 65., 67., 69., 71.,
        73., 75., 77., 79., 81.]),
 <BarContainer object of 30 artists>)
```



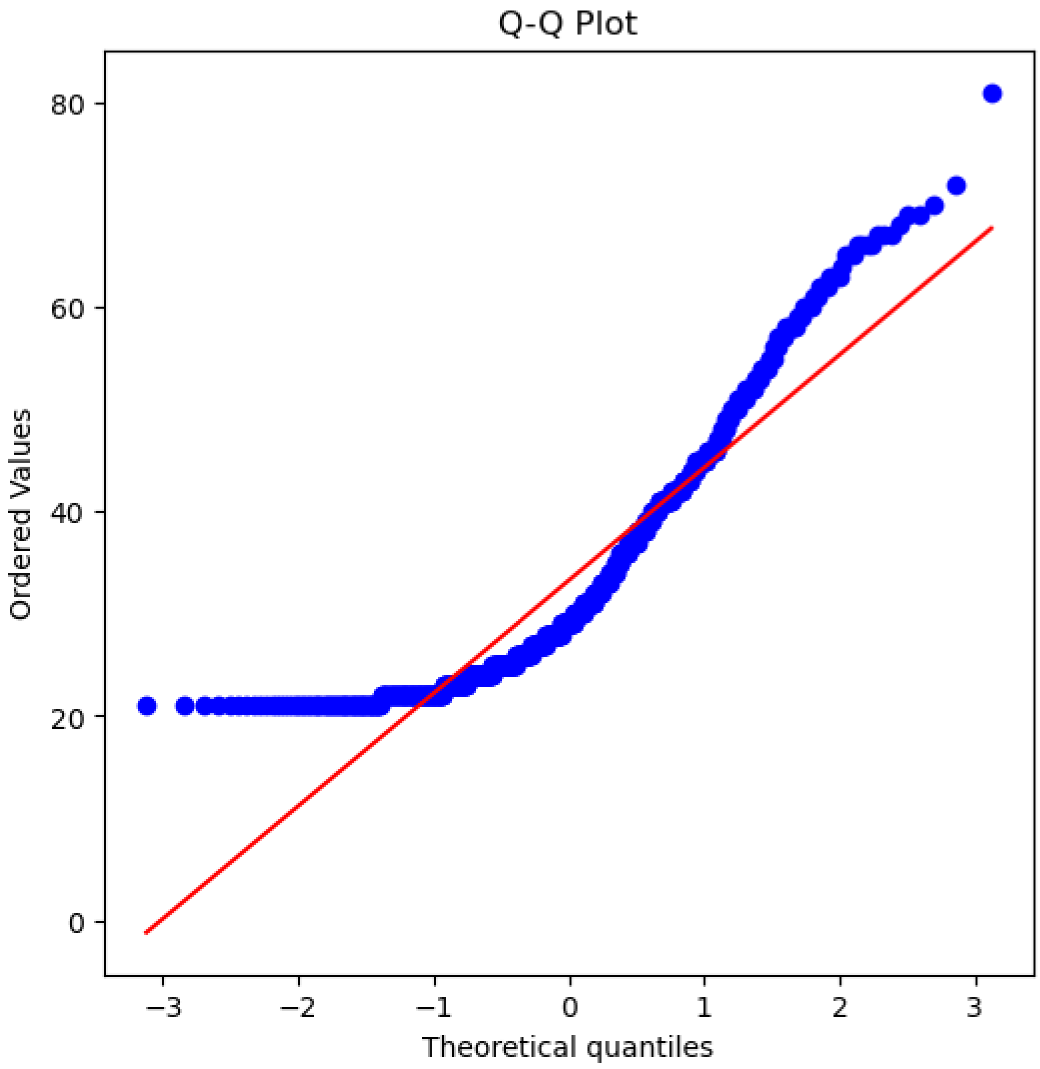
Q-Q (Quantile-Quantile) plot is a graphical tool that compares the distribution of a dataset to a theoretical

normal distribution to check if the data follows normality

In [55]:

```
import scipy.stats as stats
import matplotlib.pyplot as plt

plt.figure(figsize=(6,6))
stats.probplot(data['Age'], dist="norm", plot=plt)
plt.title("Q-Q Plot")
plt.show()
```


Q-Q Plot
Ordered Values
80
60
40
20
0
−3
−2
−1
0
1
2
3
Theoretical quantiles