Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression ) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

In [1]:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
%matplotlib inline
```

In [2]:

```
path = ("/Users/sudhanshubiswal/Project/titanic.csv")
df = pd.read_csv(path)
```

In [3]:

```
df.head()
```

Out[3]:

| | Passengerid | Age | Fare | Sex | sibsp | zero | zero.1 | zero.2 | zero.3 | zero.4 | ... | zero.12 | zerc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 22.0 | 7.2500 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| **1** | 2 | 38.0 | 71.2833 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| **2** | 3 | 26.0 | 7.9250 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| **3** | 4 | 35.0 | 53.1000 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |
| **4** | 5 | 35.0 | 8.0500 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 | |

5 rows × 28 columns

In [4]:

```
# Here survived spelling is wrong in the dataset
df. rename(columns = {'2urvived':'survived'}, inplace = True)
```

In [5]:

```
df.sample(2)
```

Out[5]:

| o.4 | ... | zero.12 | zero.13 | zero.14 | Pclass | zero.15 | zero.16 | Embarked | zero.17 | zero.18 | survived |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ... | 0 | 0 | 0 | 3 | 0 | 0 | 2.0 | 0 | 0 | 0 |
| 0 | ... | 0 | 0 | 0 | 3 | 0 | 0 | 1.0 | 0 | 0 | 1 |

# 1. Categorical data

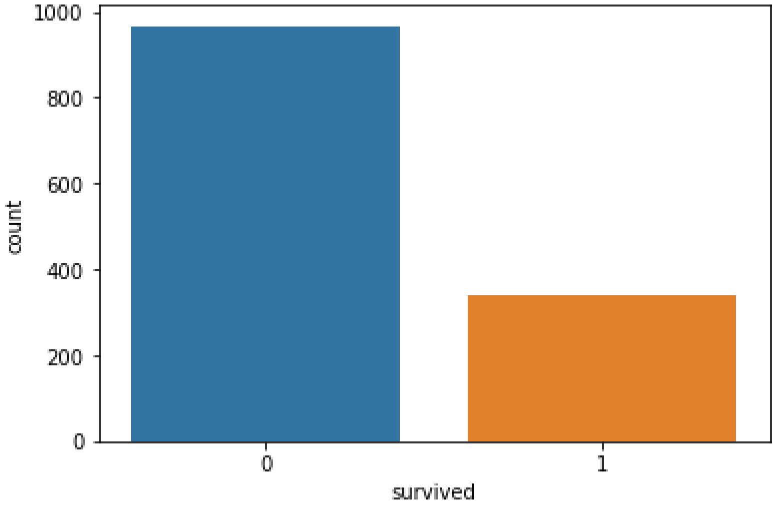
# a. Countplot

In [7]:

```
sns.countplot(df['survived'])
df['survived'].value_counts()
```

Out[7]:

```
0    967
1    342
Name: survived, dtype: int64
```



# b. Piechart

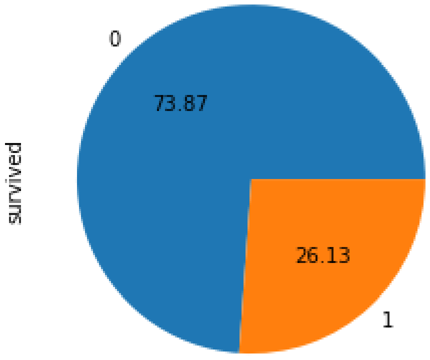
In [11]:

```
df['survived'].value_counts().plot(kind='pie',autopct='%.2f')

```

Out[11]:

```
<AxesSubplot:ylabel='survived'>
```



# 2. Numerical data

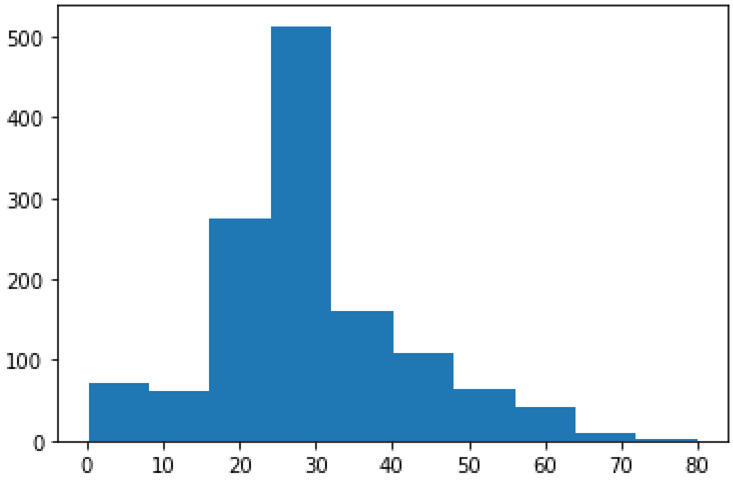
# a. Histogram

In [14]:

```
plt.hist(df['Age'])
```

Out[14]:

```
(array([ 72.,  62., 274., 513., 161., 108.,  65.,  41.,  10.,   3.]),
 array([ 0.17 ,  8.153, 16.136, 24.119, 32.102, 40.085, 48.068, 56.05
1,
        64.034, 72.017, 80.   ]),
 <BarContainer object of 10 artists>)
```



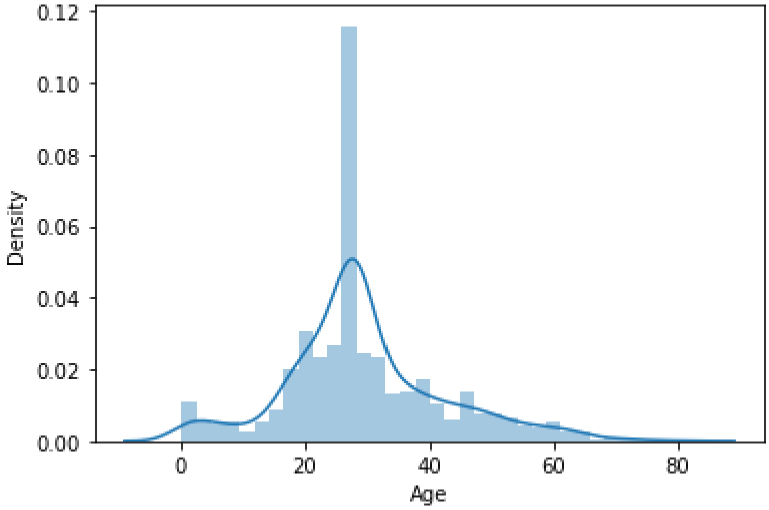
# b. Distplot

In [17]:

```
sns.distplot(df['Age'])
```

Out[17]:

```
<AxesSubplot:xlabel='Age', ylabel='Density'>
```



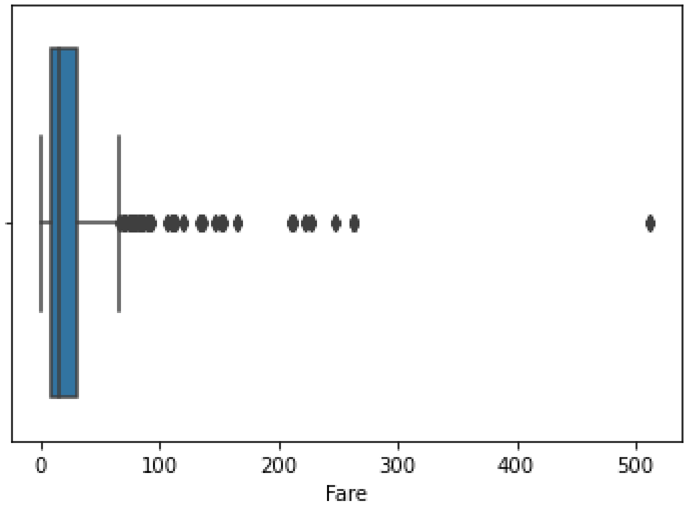
# c. Boxplot

In [18]:

```
# It is used for finding the outliers
sns.boxplot(df['Fare'])
```

Out[18]:

```
<AxesSubplot:xlabel='Fare'>
```

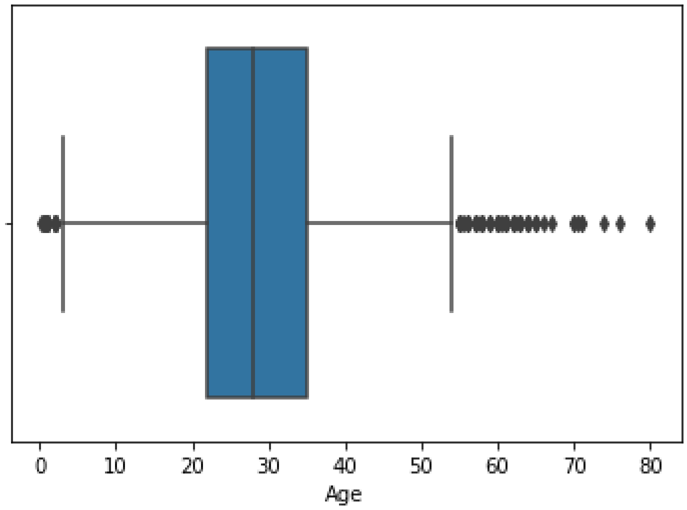


In [19]:

```
sns.boxplot(df['Age'])
```

Out[19]:

```
<AxesSubplot:xlabel='Age'>
```



In [ ]:

```

```