In [1]:
```
import pandas as pd
import numpy as np
```

In [12]:
```
df = pd.read_csv('C:\\Users\\ce\\BigDataAnalytics\\dataset\\titanic_train.csv')
```

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

```
In [4]: df.head(10)
```

Out[4]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | Na |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C8 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | Na |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C12 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | Na |
| **5** | 6 | 0 | 3 | Moran, Mr. James | male | NaN | 0 | 0 | 330877 | 8.4583 | Na |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E4 |
| **7** | 8 | 0 | 3 | Palsson, Master. Gosta Leonard | male | 2.0 | 3 | 1 | 349909 | 21.0750 | Na |
| **8** | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27.0 | 0 | 2 | 347742 | 11.1333 | Na |
| **9** | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14.0 | 1 | 0 | 237736 | 30.0708 | Na |

In [5]:

```
df.dropna(how='any')
```

Out[5]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cab |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C1 |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E |
| **10** | 11 | 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 4.0 | 1 | 1 | PP 9549 | 16.7000 | G |
| **11** | 12 | 1 | 1 | Bonnell, Miss. Elizabeth | female | 58.0 | 0 | 0 | 113783 | 26.5500 | C1 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **871** | 872 | 1 | 1 | Beckwith, Mrs. Richard Leonard (Sallie Monypeny) | female | 47.0 | 1 | 1 | 11751 | 52.5542 | D |
| **872** | 873 | 0 | 1 | Carlsson, Mr. Frans Olof | male | 33.0 | 0 | 0 | 695 | 5.0000 | B<br>B<br>B |
| **879** | 880 | 1 | 1 | Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) | female | 56.0 | 0 | 1 | 11767 | 83.1583 | C |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 | 0 | 112053 | 30.0000 | B |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.0000 | C1 |

183 rows × 12 columns

In [7]:
```
df.shape
```

Out[7]:
```
(891, 12)
```

In [8]:
```
df.dropna(subset=['Age'])
```

Out[8]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Ca |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | I |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | I |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | I |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **885** | 886 | 0 | 3 | Rice, Mrs. William (Margaret Norton) | female | 39.0 | 0 | 5 | 382652 | 29.1250 | I |
| **886** | 887 | 0 | 2 | Montvila, Rev. Juozas | male | 27.0 | 0 | 0 | 211536 | 13.0000 | I |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 | 0 | 112053 | 30.0000 | |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.0000 | C |
| **890** | 891 | 0 | 3 | Dooley, Mr. Patrick | male | 32.0 | 0 | 0 | 370376 | 7.7500 | I |

714 rows × 12 columns

In [9]:
```
df.shape
```

Out[9]:
```
(891, 12)
```

```
In [10]: df.dropna(subset=['Age'],inplace=True)
```

```
In [16]: df.shape
```

```
Out[16]: (891, 12)
```

```
In [13]: df.shape
```

```
Out[13]: (891, 12)
```

In [18]:

```
df.dropna(subset=['Cabin','Age'])
```

Out[18]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cab |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C8 |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C12 |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E4 |
| **10** | 11 | 1 | 3 | Sandstrom, Miss. Marguerite Rut | female | 4.0 | 1 | 1 | PP 9549 | 16.7000 | G |
| **11** | 12 | 1 | 1 | Bonnell, Miss. Elizabeth | female | 58.0 | 0 | 0 | 113783 | 26.5500 | C10 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **871** | 872 | 1 | 1 | Beckwith, Mrs. Richard Leonard (Sallie Monypeny) | female | 47.0 | 1 | 1 | 11751 | 52.5542 | D3 |
| **872** | 873 | 0 | 1 | Carlsson, Mr. Frans Olof | male | 33.0 | 0 | 0 | 695 | 5.0000 | B5<br>B5<br>B5 |
| **879** | 880 | 1 | 1 | Potter, Mrs. Thomas Jr (Lily Alexenia Wilson) | female | 56.0 | 0 | 1 | 11767 | 83.1583 | C5 |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 | 0 | 112053 | 30.0000 | B4 |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.0000 | C14 |

185 rows × 12 columns

In [19]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          714 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [20]:

```
df.drop(['Cabin'],axis=1)
```

Out[20]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | E |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| **886** | 887 | 0 | 2 | Montvila, Rev. Juozas | male | 27.0 | 0 | 0 | 211536 | 13.0000 | |
| **887** | 888 | 1 | 1 | Graham, Miss. Margaret Edith | female | 19.0 | 0 | 0 | 112053 | 30.0000 | |
| **888** | 889 | 0 | 3 | Johnston, Miss. Catherine Helen "Carrie" | female | NaN | 1 | 2 | W./C. 6607 | 23.4500 | |
| **889** | 890 | 1 | 1 | Behr, Mr. Karl Howell | male | 26.0 | 0 | 0 | 111369 | 30.0000 | |
| **890** | 891 | 0 | 3 | Dooley, Mr. Patrick | male | 32.0 | 0 | 0 | 370376 | 7.7500 | |

891 rows × 11 columns

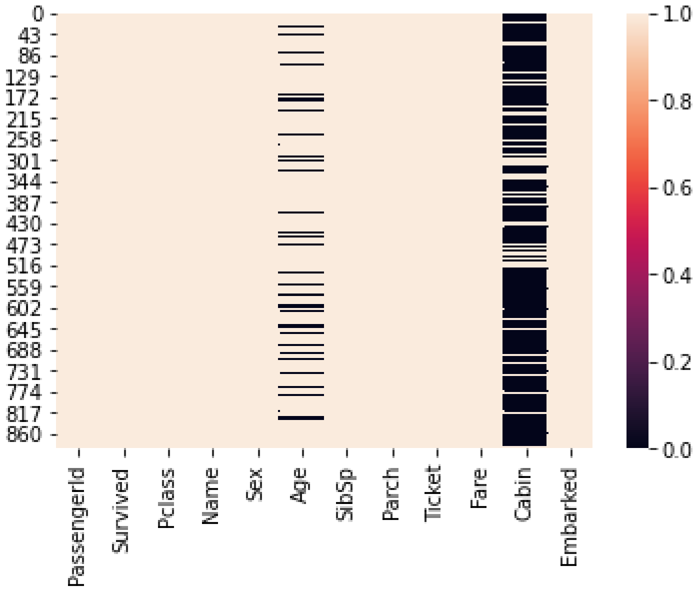
In [27]:
```
df.head(10)
```

Out[27]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | Na |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C8 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | Na |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C12 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | Na |
| **5** | 6 | 0 | 3 | Moran, Mr. James | male | NaN | 0 | 0 | 330877 | 8.4583 | Na |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.0 | 0 | 0 | 17463 | 51.8625 | E4 |
| **7** | 8 | 0 | 3 | Palsson, Master. Gosta Leonard | male | 2.0 | 3 | 1 | 349909 | 21.0750 | Na |
| **8** | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27.0 | 0 | 2 | 347742 | 11.1333 | Na |
| **9** | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14.0 | 1 | 0 | 237736 | 30.0708 | Na |

In [23]:
```
import seaborn as sns
sns.heatmap(df.notnull())
```

Out[23]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2716938a190>
```



In [31]:
```
df['Age'].fillna(df['Age'].mean(),inplace = True)
```

In [32]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   PassengerId  891 non-null    int64  
 1   Survived     891 non-null    int64  
 2   Pclass       891 non-null    int64  
 3   Name         891 non-null    object 
 4   Sex          891 non-null    object 
 5   Age          891 non-null    float64
 6   SibSp        891 non-null    int64  
 7   Parch        891 non-null    int64  
 8   Ticket       891 non-null    object 
 9   Fare         891 non-null    float64
 10  Cabin        204 non-null    object 
 11  Embarked     889 non-null    object 
dtypes: float64(2), int64(5), object(5)
memory usage: 83.7+ KB
```

In [33]:
```
df.head(10)
```

Out[33]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.000000 | 1 | 0 | A/5 21171 | 7.2500 |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.000000 | 1 | 0 | PC 17599 | 71.2833 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.000000 | 0 | 0 | STON/O2. 3101282 | 7.9250 |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.000000 | 1 | 0 | 113803 | 53.1000 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.000000 | 0 | 0 | 373450 | 8.0500 |
| **5** | 6 | 0 | 3 | Moran, Mr. James | male | 29.699118 | 0 | 0 | 330877 | 8.4583 |
| **6** | 7 | 0 | 1 | McCarthy, Mr. Timothy J | male | 54.000000 | 0 | 0 | 17463 | 51.8625 |
| **7** | 8 | 0 | 3 | Palsson, Master. Gosta Leonard | male | 2.000000 | 3 | 1 | 349909 | 21.0750 |
| **8** | 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | female | 27.000000 | 0 | 2 | 347742 | 11.1333 |
| **9** | 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | female | 14.000000 | 1 | 0 | 237736 | 30.0708 |

# --- 02-22-2021

In [2]:
```
import numpy as np
import pandas as pd
```

In [3]:
```
import seaborn as sns
```

In [4]:
```
df = pd.read_csv('C:\\Users\\ce\\BigDataAnalytics\\dataset\\titanic_train.csv')
```

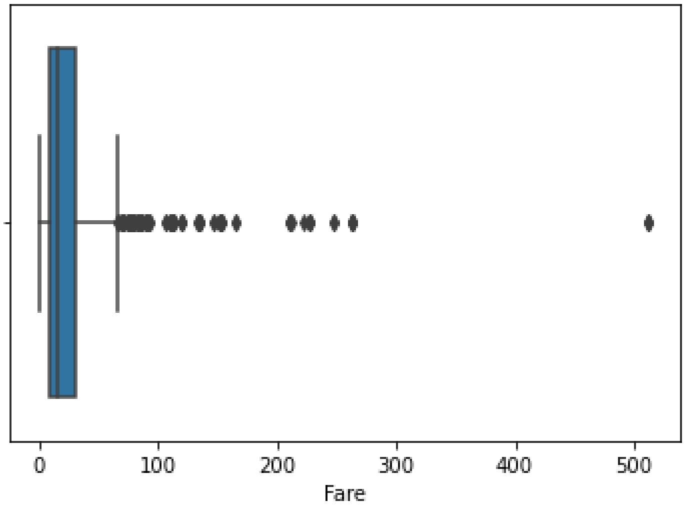
In [5]:
```
df.head()
```

Out[5]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | Na |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C8 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | Na |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C12 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | Na |

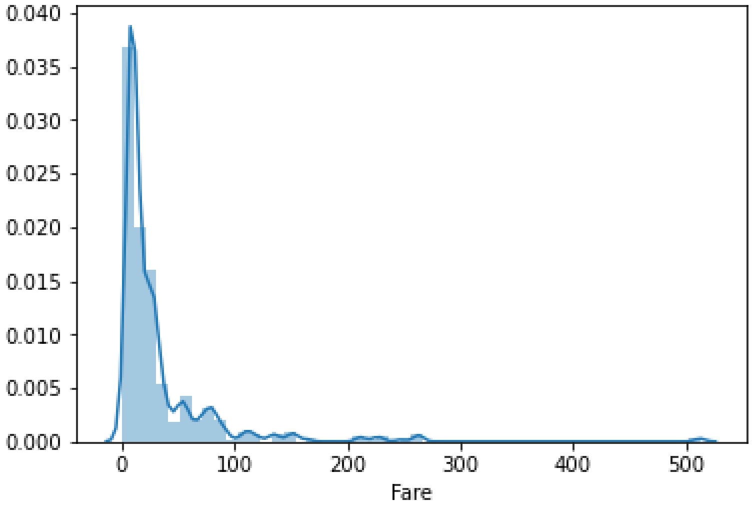
In [6]:
```
sns.boxplot(df['Fare'])
```

Out[6]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2074e500a90>
```



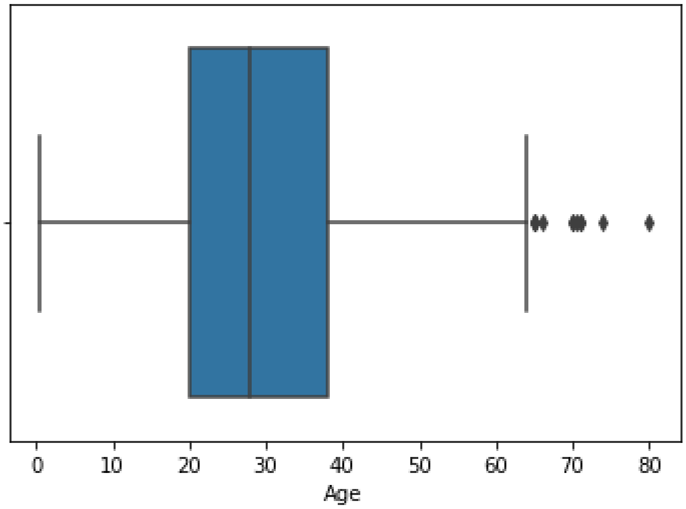
In [7]:
```
sns.distplot(df['Fare'])
```

Out[7]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2074e5d6820>
```



In [8]:
```
sns.boxplot(df['Age'])
```

Out[8]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2074e6cc490>
```

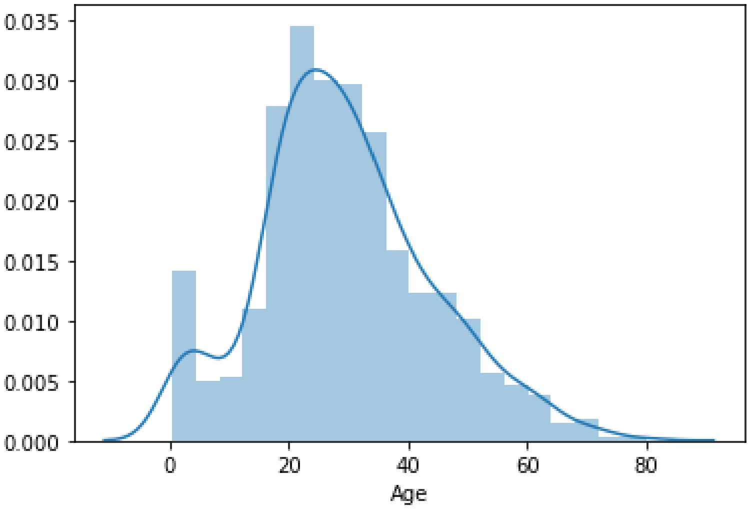


In [9]:

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

Out[9]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x2074e729640>
```



In [1]:

```
import pandas as pd
import numpy as np
```

In [10]:

```
dfW = pd.read_csv('D:\\Teaching Subject\\Data Science\\Fall 2021\\Lectures\\Struc
```

In [11]:

```
dfW
```

Out[11]:

| | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 # | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34FA | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

In [12]:
```
dfW['temperature'].replace('[^0-9-]','',inplace=True,regex=True)
```

In [14]:
```
dfW
```

Out[14]:

| | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34 | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

# --- 09-11-2021

In [15]:
```
def find_boundaries(df, variable, distance):
    IQR = df[variable].quantile(0.75) - df[variable].quantile(0.25)
    lower_boundary = df[variable].quantile(0.25) - (IQR * distance)
    upper_boundary = df[variable].quantile(0.75) + (IQR * distance)
    return upper_boundary, lower_boundary
```

In [36]:
```
df = pd.read_csv('C:\\Users\\ce\\BigDataAnalytics\\dataset\\titanic_train.csv')
```

In [30]:
```
df.shape
```

Out[30]:
```
(1309, 10)
```

In [19]:
```
df.head()
```

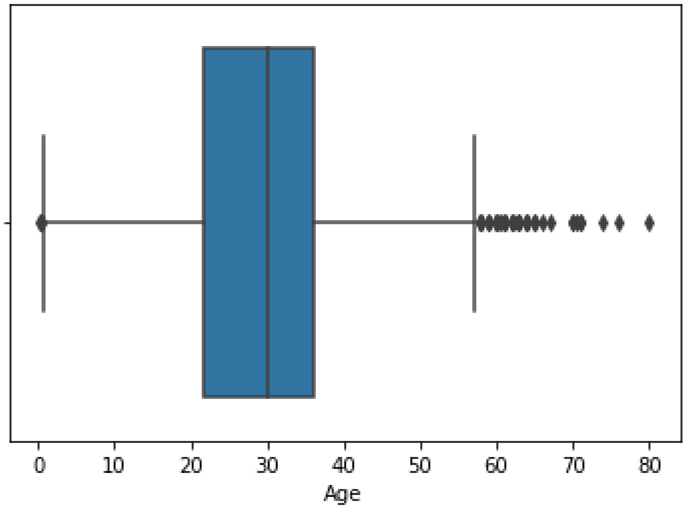
Out[19]:

| | Unnamed: 0 | Pclass | Age | SibSp | Parch | Fare | Embarked | Title | Gen_male | Survived |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 3 | 22.0 | 1 | 0 | 7.2500 | 0 | 0 | 1 | 0 |
| **1** | 1 | 1 | 38.0 | 1 | 0 | 71.2833 | 1 | 1 | 0 | 1 |
| **2** | 2 | 3 | 26.0 | 0 | 0 | 7.9250 | 0 | 2 | 0 | 1 |
| **3** | 3 | 1 | 35.0 | 1 | 0 | 53.1000 | 0 | 1 | 0 | 1 |
| **4** | 4 | 3 | 35.0 | 0 | 0 | 8.0500 | 0 | 0 | 1 | 0 |

In [20]:
```
import seaborn as sns
```

```
In [21]: sns.boxplot(df['Age'])
```

```
Out[21]: <matplotlib.axes._subplots.AxesSubplot at 0x2de5eaef9d0>
```



```
In [23]: upL,LwL = find_boundaries(df, 'Age', 1.5)
```

```
In [25]: outL_df = np.where(df['Age'] > upL, True, np.where(df['Age'] < LwL, True,
         False))
```

```
In [26]: outL_df
```

```
Out[26]: array([False, False, False, ..., False, False, False])
```

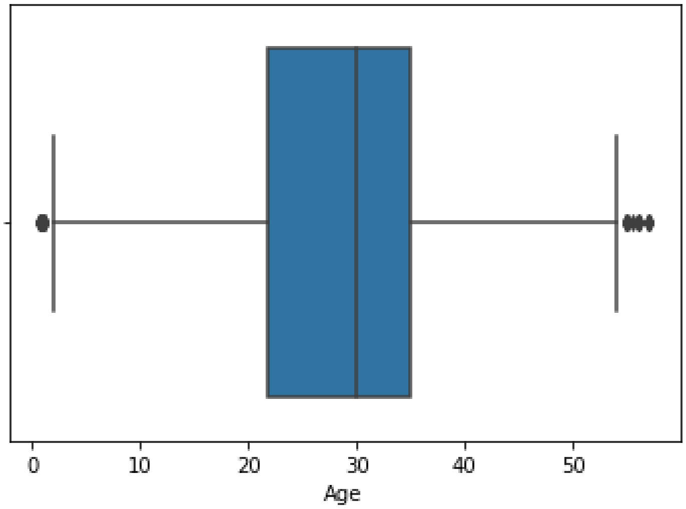
```
In [27]: df_new = df.loc[~(outL_df)]
```

```
In [29]: df_new.shape
```

```
Out[29]: (1257, 10)
```

In [31]:
```
sns.boxplot(df_new['Age'])
```

Out[31]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x2de5ebc2040>
```



In [33]:
```
find_boundaries(df_new,'Age',1.7)
```

Out[33]:
```
(57.48379523, -0.7095571299999968)
```

In [34]:

```
df_new
```

Out[34]:

```
      Unnamed:  Pclass        Age  SibSp  Parch      Fare  Embarked  Title  Gen_male  Survived
             0
   0         0       3  22.000000      1      0    7.2500         0      0         1         0
   1         1       1  38.000000      1      0   71.2833         1      1         0         1
   2         2       3  26.000000      0      0    7.9250         0      2         0         1
   3         3       1  35.000000      1      0   53.1000         0      1         0         1
   4         4       3  35.000000      0      0    8.0500         0      0         1         0
 ...       ...     ...        ...    ...    ...       ...       ...    ...       ...       ...
1304       413       3  32.252151      0      0    8.0500         0      0         1         1
1305       414       1  39.000000      0      0  108.9000         1      1         0         0
1306       415       3  38.500000      0      0    7.2500         0      0         1         1
1307       416       3  32.252151      0      0    8.0500         0      0         1         1
1308       417       3   5.482642      1      1   22.3583         1      1         1         0

1257 rows × 10 columns
```

In [37]:

```
df['Title'] = df['Name'].str.extract('([A-Za-z]+\.)',expand=False)
```

In [38]:

```
df.head()
```

Out[38]:

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | Na |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C8 |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | Na |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C12 |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | Na |

In [39]:

```
df['Title'].value_counts()
```

Out[39]:

```
Mr.          517
Miss.        182
Mrs.         125
Master.       40
Dr.            7
Rev.           6
Mlle.          2
Col.           2
Major.         2
Capt.          1
Ms.            1
Mme.           1
Countess.      1
Don.           1
Sir.           1
Lady.          1
Jonkheer.      1
Name: Title, dtype: int64
```

In [1]:

```
import pandas as pd
```

In [2]:

```
dfW = pd.read_csv('D:\\Teaching Subject\\Data Science\\Fall 2021\\Lectures\\Struc
```

In [3]:

```
dfW
```

Out[3]:

|   | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 # | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34FA | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

In [6]:

```
import re
```

In [9]:

```
re.findall('[-]?[0-9]+', str(dfW['temperature']))
```

Out[9]:

```
['0',
 '32',
 '1',
 '-9999',
 '2',
 '28',
 '3',
 '-9999',
 '4',
 '32',
 '5',
 '-9999',
 '6',
 '-9999',
 '7',
 '34',
 '8',
 '40']
```

In [ ]:

```
dfW['temperature'].replace('[-]?[0-9]+','',inplace=True,regex=True)
```

In [5]:
```
dfW
```

Out[5]:

| | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 # | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34FA | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

In [10]:
```
dfW
```

Out[10]:

| | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 # | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34FA | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

In [11]:
```
dfW['temperature'].replace('[^0-9-]','',inplace=True,regex=True)
```

In [12]:

```
dfW
```

Out[12]:

|  | day | temperature | windspeed | event |
|---|---|---|---|---|
| **0** | 1/1/2017 | 32 | 6us | Rain |
| **1** | 1/4/2017 | -9999 | 9 | Sunny |
| **2** | 1/5/2017 | 28 | -7777 | Snow |
| **3** | 1/6/2017 | -9999 | 7 | NaN |
| **4** | 1/7/2017 | 32 | -7777 | Rain |
| **5** | 1/8/2017 | -9999 | -7777 | Sunny |
| **6** | 1/9/2017 | -9999 | -7777 | NaN |
| **7** | 1/10/2017 | 34 | 8yyy | Cloudy |
| **8** | 1/11/2017 | 40 | 12 | Sunny |

In [ ]:

```

```