In [1]:

```
#1)Load the libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

In [2]:

```
#2) Download the data set from kaggle/ other sources
```

In [3]:

```
#3) Read the file -select appropriate file read function according to data type of file
df = pd.read_csv("heart.csv")
```

In [4]:

```
#4) Display attributes in the data set-10 samples.
df.head(10)
```

Out[4]:

| | Age | Sex | ChestPain | RestBP | Chol | Fbs | RestECG | MaxHR | ExAng | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 63 | 1 | typical | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 |
| **1** | 67 | 1 | asymptomatic | 160 | 286 | 0 | 2 | 108 | 1 | 1.5 | 2 |
| **2** | 67 | 1 | asymptomatic | 120 | 229 | 0 | 2 | 129 | 1 | 2.6 | 2 |
| **3** | 37 | 1 | nonanginal | 130 | 250 | 0 | 0 | 187 | 0 | 3.5 | 3 |
| **4** | 41 | 0 | nontypical | 130 | 204 | 0 | 2 | 172 | 0 | 1.4 | 1 |
| **5** | 56 | 1 | nontypical | 120 | 236 | 0 | 0 | 178 | 0 | 0.8 | 1 |
| **6** | 62 | 0 | asymptomatic | 140 | 268 | 0 | 2 | 160 | 0 | 3.6 | 3 |
| **7** | 57 | 0 | asymptomatic | 120 | 354 | 0 | 0 | 163 | 1 | 0.6 | 1 |
| **8** | 63 | 1 | asymptomatic | 130 | 254 | 0 | 2 | 147 | 0 | 1.4 | 2 |
| **9** | 53 | 1 | asymptomatic | 140 | 203 | 1 | 2 | 155 | 1 | 3.1 | 3 |

In [5]:

```
#5) Describe the attributes find  range, quartile,
#percentile, box plot and outliers.
df.describe()
```

Out[5]:

| | Age | Sex | RestBP | Chol | Fbs | RestECG | MaxHR |
|---|---|---|---|---|---|---|---|
| **count** | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 | 303.000000 |
| **mean** | 54.438944 | 0.679868 | 131.689769 | 246.693069 | 0.148515 | 0.990099 | 149.607261 |
| **std** | 9.038662 | 0.467299 | 17.599748 | 51.776918 | 0.356198 | 0.994971 | 22.875003 |
| **min** | 29.000000 | 0.000000 | 94.000000 | 126.000000 | 0.000000 | 0.000000 | 71.000000 |
| **25%** | 48.000000 | 0.000000 | 120.000000 | 211.000000 | 0.000000 | 0.000000 | 133.500000 |
| **50%** | 56.000000 | 1.000000 | 130.000000 | 241.000000 | 0.000000 | 1.000000 | 153.000000 |
| **75%** | 61.000000 | 1.000000 | 140.000000 | 275.000000 | 0.000000 | 2.000000 | 166.000000 |
| **max** | 77.000000 | 1.000000 | 200.000000 | 564.000000 | 1.000000 | 2.000000 | 202.000000 |

In [12]:

```
#5) Describe the attributes name, data type
df.dtypes
```

Out[12]:

```
Age            int64
Sex            int64
ChestPain     object
RestBP         int64
Chol           int64
Fbs            int64
RestECG        int64
MaxHR          int64
ExAng          int64
Oldpeak      float64
Slope          int64
Ca           float64
Thal          object
AHD           object
dtype: object
```

In [8]:

```
print('Q1 : 25th percentile of arr :', np.percentile(df['RestBP'], 25))
print('Q2 : 50th percentile of arr :', np.percentile(df['RestBP'], 50))
print('Q3 : 75th percentile of arr :', np.percentile(df['RestBP'], 75))
```

```
Q1 : 25th percentile of arr : 120.0
Q2 : 50th percentile of arr : 130.0
Q3 : 75th percentile of arr : 140.0
```

In [9]:

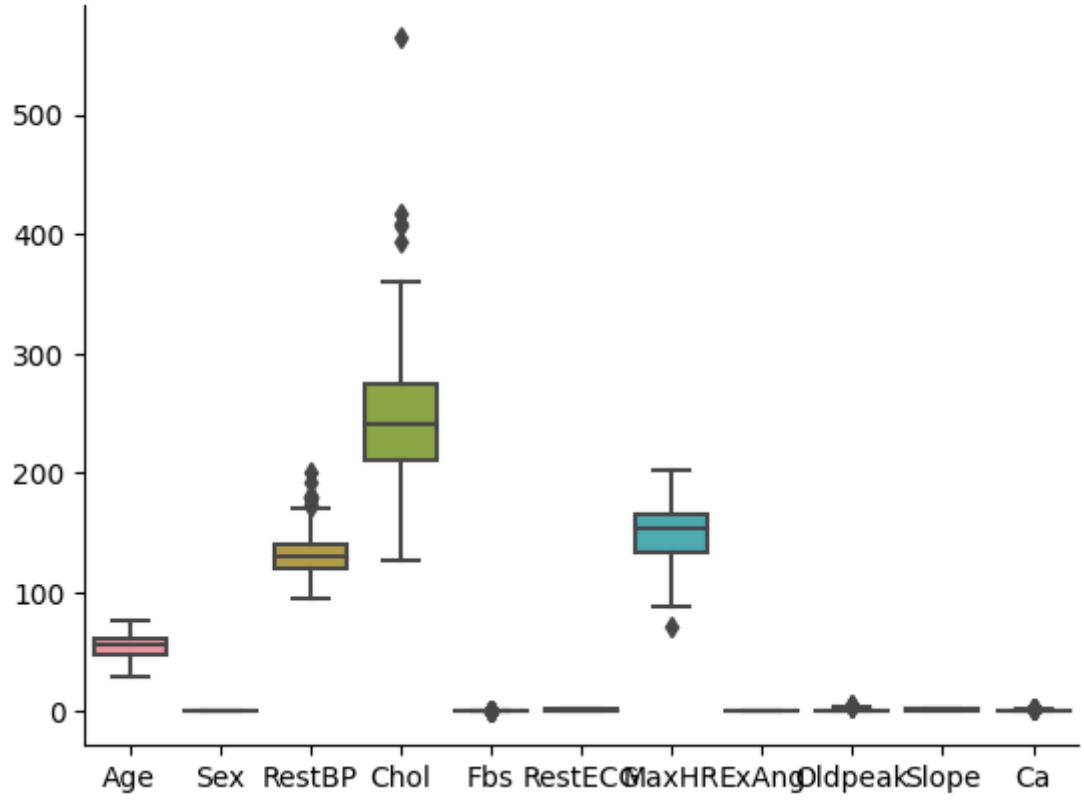
```
np.quantile(df['RestBP'], [0,0.25,0.5,0.75,1])
```

Out[9]:

```
array([ 94., 120., 130., 140., 200.])
```

In [11]:

```
sns.boxplot(df)
sns.despine()
# all dots lying above or below box plots are outliers
```



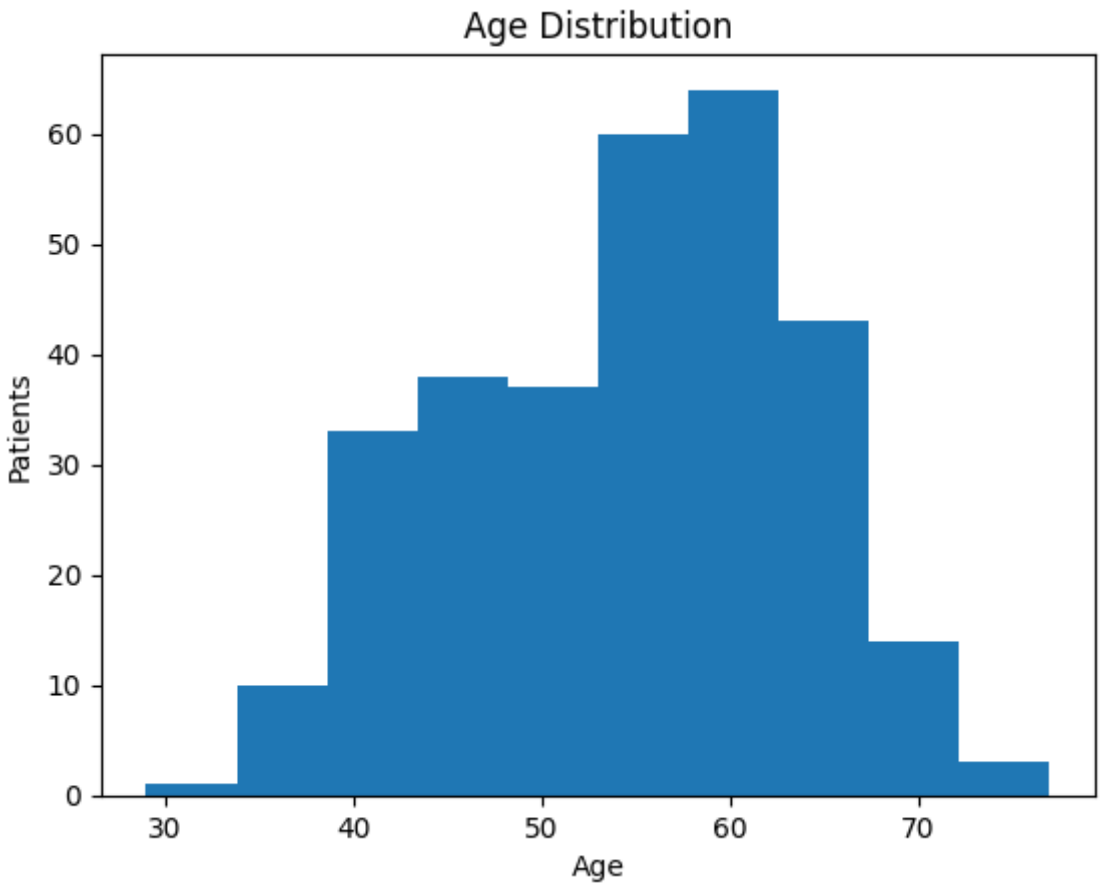
In [13]:

```
## HISTOGRAM ##
fig=plt.figure()
ax = fig.add_subplot(1,1,1)

#Variable
ax.hist(df['Age'],bins = 10)
#Labels and Tit
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Patients')
```

Out[13]:

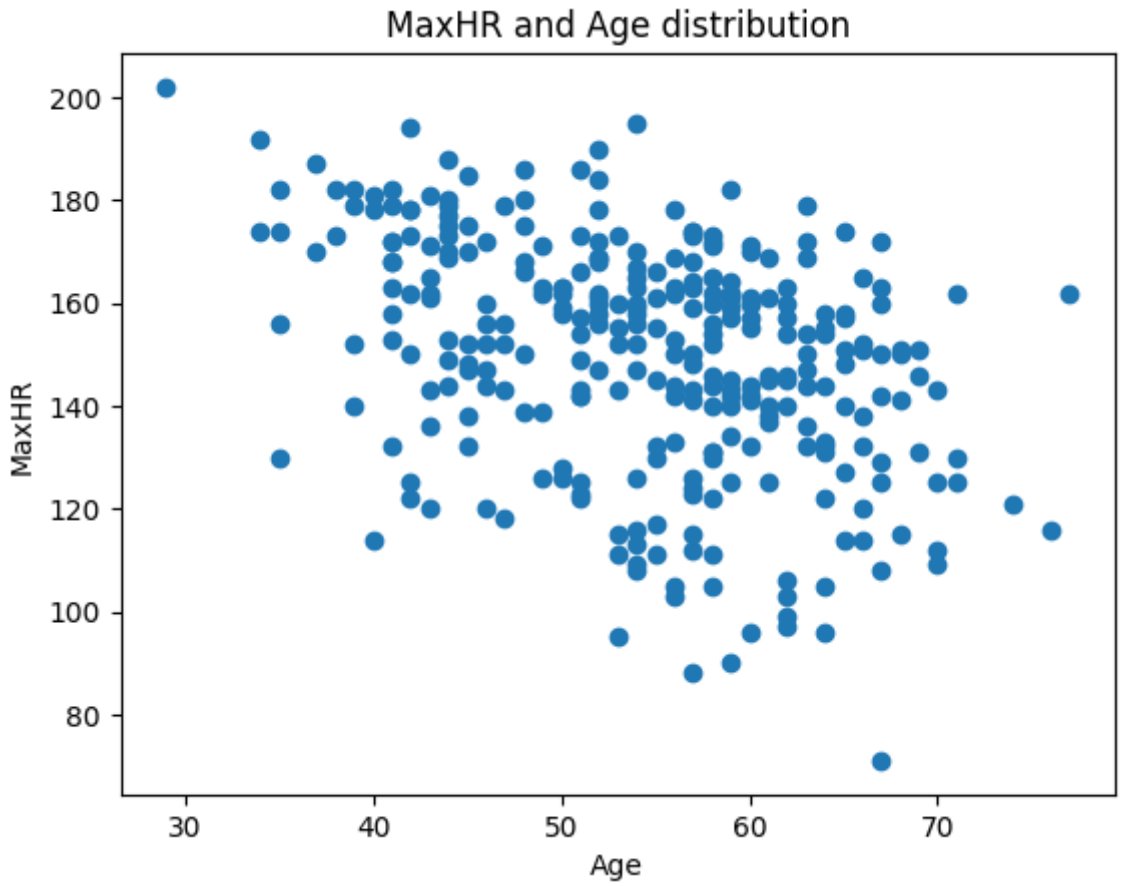
```
Text(0, 0.5, 'Patients')
```



In [14]:

```
## SCATTER PLOT ##
fig=plt.figure()
ax = fig.add_subplot(1,1,1)

#Variable
ax.scatter(df['Age'],df['MaxHR'])
#Labels and Tit
plt.title('MaxHR and Age distribution')
plt.xlabel('Age')
plt.ylabel('MaxHR')
plt.show()
```




In [19]:

```
## FREQUENCY TABLE ##
test= df.groupby(['Sex','ChestPain'])
test.size()
```

Out[19]:

```
Sex  ChestPain
0    asymptomatic     40
     nonanginal       35
     nontypical       18
     typical           4
1    asymptomatic    104
     nonanginal       51
     nontypical       32
     typical          19
dtype: int64
```

In [20]:

```
#7) Give correlation matrix
matrix = np.corrcoef(df['Age'], df['RestBP'])
print(matrix)
```

```
[[1.         0.28494592]
 [0.28494592 1.        ]]
```

In [21]:

```
#8) Identify missing values and outlier and fill them with average.
# to check all missing values
df.isna().sum()
```

Out[21]:

```
Age         0
Sex         0
ChestPain   0
RestBP      0
Chol        0
Fbs         0
RestECG     0
MaxHR       0
ExAng       0
Oldpeak     0
Slope       0
Ca          4
Thal        2
AHD         0
dtype: int64
```

In [22]:

```
# replacing missing values in the DataFrame
meanCa = np.mean(df.Ca)
df.Ca = df.Ca.fillna(meanCa)
```

In [24]:

```
# to remove missing values
df = df.dropna()
```

In [25]:

```
df.isna().sum()
```

Out[25]:

```
Age          0
Sex          0
ChestPain    0
RestBP       0
Chol         0
Fbs          0
RestECG      0
MaxHR        0
ExAng        0
Oldpeak      0
Slope        0
Ca           0
Thal         0
AHD          0
dtype: int64
```