In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

In [8]:

```
df = pd.read_csv(r"C:\Users\Vinayak\Downloads\dsbda\B_ass2\heart_disease\heart.csv",sep=
df.head()
```

Out[8]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	60	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	35	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	55	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	56	0	0	120	354	0	1	163	1	0.6	2	0	2	1

In [5]:

```
df.describe()
```

Out[5]:

	age	sex	cp	trtbps	chol	fbs	restecg
count	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000	289.000000
mean	54.010381	0.678201	1.020761	131.377163	247.961938	0.145329	0.515571
std	9.132316	0.467977	1.027192	17.518432	51.596208	0.353043	0.514309
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000
25%	47.000000	0.000000	0.000000	120.000000	212.000000	0.000000	0.000000
50%	54.000000	1.000000	1.000000	130.000000	243.000000	0.000000	1.000000
75%	60.000000	1.000000	2.000000	140.000000	276.000000	0.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000

In [ ]:

```
##data cleaning
```

In [6]:

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

Out[6]:

```
age         0
sex         0
cp          0
trtbps      0
chol        0
fbs         0
restecg     0
thalachh    0
exng        0
oldpeak     0
slp         0
caa         0
thall       0
output      0
dtype: int64
```

In [9]:

```
#removing unwanted columns
select = ['age', 'sex', 'cp', 'trtbps', 'chol']
df1 = df[select].copy()
df1
```

Out[9]:

	age	sex	cp	trtbps	chol
0	60	1	3	145	233
1	35	1	2	130	250
2	41	0	1	130	204
3	55	1	1	120	236
4	56	0	0	120	354
5	55	1	0	140	192
6	56	0	1	140	294
7	44	1	1	120	263
8	52	1	2	172	199
9	57	1	2	150	168
10	54	1	0	140	239
11	48	0	2	130	275
12	49	1	1	130	266
13	64	1	3	110	211
14	55	0	3	150	283
15	50	0	2	120	219
16	58	0	2	120	340
17	66	0	3	150	226
18	40	1	0	150	247
19	69	0	3	140	239
20	59	1	0	135	234
21	44	1	2	130	233
22	42	1	0	140	226
23	61	1	2	150	243
24	40	1	3	140	199
25	71	0	1	160	302
26	59	1	2	150	212
27	51	1	2	110	175
28	65	0	2	140	417
29	53	1	2	130	197
...	...	...	...	...	...
259	38	1	3	120	231
260	66	0	0	178	228
261	52	1	0	112	230
262	53	1	0	123	282
263	63	0	0	108	269
264	54	1	0	110	206

	age	sex	cp	trtbps	chol
265	66	1	0	112	212
266	55	0	0	180	327
267	49	1	2	118	149
268	54	1	0	122	286
269	52	1	0	130	283
270	46	1	0	120	249
271	60	1	3	134	234
272	67	1	0	120	237
273	58	1	0	100	234
274	45	1	0	110	275
275	52	1	0	125	212
276	58	1	0	146	218
277	55	1	1	124	261
278	58	0	1	136	319
279	61	1	0	138	166
280	42	1	0	136	315
281	50	1	0	128	204
282	59	1	2	126	218
283	40	1	0	152	223
284	60	1	0	140	207
285	46	1	0	140	311
286	59	1	3	134	204
287	54	1	1	154	232
288	53	1	0	110	335

289 rows × 5 columns

In [ ]:

```
#data integration
```

In [11]:

```
sub1 =df[['age','sex']]
sub1
```

Out[11]:

	age	sex
0	60	1
1	35	1
2	41	0
3	55	1
4	56	0
5	55	1
6	56	0
7	44	1
8	52	1
9	57	1
10	54	1
11	48	0
12	49	1
13	64	1
14	55	0
15	50	0
16	58	0
17	66	0
18	40	1
19	69	0
20	59	1
21	44	1
22	42	1
23	61	1
24	40	1
25	71	0
26	59	1
27	51	1
28	65	0
29	53	1
...	...	...
259	38	1
260	66	0
261	52	1
262	53	1
263	63	0
264	54	1

	age	sex
265	66	1
266	55	0
267	49	1
268	54	1
269	52	1
270	46	1
271	60	1
272	67	1
273	58	1
274	45	1
275	52	1
276	58	1
277	55	1
278	58	0
279	61	1
280	42	1
281	50	1
282	59	1
283	40	1
284	60	1
285	46	1
286	59	1
287	54	1
288	53	1

289 rows × 2 columns

In [12]:

```
sub2 =df[['cp','chol']]
sub2
```

Out[12]:

	cp	chol
0	3	233
1	2	250
2	1	204
3	1	236
4	0	354
5	0	192
6	1	294
7	1	263
8	2	199
9	2	168
10	0	239
11	2	275
12	1	266
13	3	211
14	3	283
15	2	219
16	2	340
17	3	226
18	0	247
19	3	239
20	0	234
21	2	233
22	0	226
23	2	243
24	3	199
25	1	302
26	2	212
27	2	175
28	2	417
29	2	197
...	...	...
259	3	231
260	0	228
261	0	230
262	0	282
263	0	269
264	0	206

	cp	chol
265	0	212
266	0	327
267	2	149
268	0	286
269	0	283
270	0	249
271	3	234
272	0	237
273	0	234
274	0	275
275	0	212
276	0	218
277	1	261
278	1	319
279	0	166
280	0	315
281	0	204
282	2	218
283	0	223
284	0	207
285	0	311
286	3	204
287	1	232
288	0	335

289 rows × 2 columns

In [13]:

```
merge = pd.concat([sub1, sub2],sort=False)
merge
```

Out[13]:

	age	sex	cp	chol
0	60.0	1.0	NaN	NaN
1	35.0	1.0	NaN	NaN
2	41.0	0.0	NaN	NaN
3	55.0	1.0	NaN	NaN
4	56.0	0.0	NaN	NaN
5	55.0	1.0	NaN	NaN
6	56.0	0.0	NaN	NaN
7	44.0	1.0	NaN	NaN
8	52.0	1.0	NaN	NaN
9	57.0	1.0	NaN	NaN
10	54.0	1.0	NaN	NaN
11	48.0	0.0	NaN	NaN
12	49.0	1.0	NaN	NaN
13	64.0	1.0	NaN	NaN
14	55.0	0.0	NaN	NaN
15	50.0	0.0	NaN	NaN
16	58.0	0.0	NaN	NaN
17	66.0	0.0	NaN	NaN
18	40.0	1.0	NaN	NaN
19	69.0	0.0	NaN	NaN
20	59.0	1.0	NaN	NaN
21	44.0	1.0	NaN	NaN
22	42.0	1.0	NaN	NaN
23	61.0	1.0	NaN	NaN
24	40.0	1.0	NaN	NaN
25	71.0	0.0	NaN	NaN
26	59.0	1.0	NaN	NaN
27	51.0	1.0	NaN	NaN
28	65.0	0.0	NaN	NaN
29	53.0	1.0	NaN	NaN
...	...	...	...	...
259	NaN	NaN	3.0	231.0
260	NaN	NaN	0.0	228.0
261	NaN	NaN	0.0	230.0
262	NaN	NaN	0.0	282.0
263	NaN	NaN	0.0	269.0
264	NaN	NaN	0.0	206.0

	age	sex	cp	chol
265	NaN	NaN	0.0	212.0
266	NaN	NaN	0.0	327.0
267	NaN	NaN	2.0	149.0
268	NaN	NaN	0.0	286.0
269	NaN	NaN	0.0	283.0
270	NaN	NaN	0.0	249.0
271	NaN	NaN	3.0	234.0
272	NaN	NaN	0.0	237.0
273	NaN	NaN	0.0	234.0
274	NaN	NaN	0.0	275.0
275	NaN	NaN	0.0	212.0
276	NaN	NaN	0.0	218.0
277	NaN	NaN	1.0	261.0
278	NaN	NaN	1.0	319.0
279	NaN	NaN	0.0	166.0
280	NaN	NaN	0.0	315.0
281	NaN	NaN	0.0	204.0
282	NaN	NaN	2.0	218.0
283	NaN	NaN	0.0	223.0
284	NaN	NaN	0.0	207.0
285	NaN	NaN	0.0	311.0
286	NaN	NaN	3.0	204.0
287	NaN	NaN	1.0	232.0
288	NaN	NaN	0.0	335.0

578 rows × 4 columns

In [ ]:

```
#Data transformation
```

In [ ]:

```
# Error Correcting
```

In [ ]:

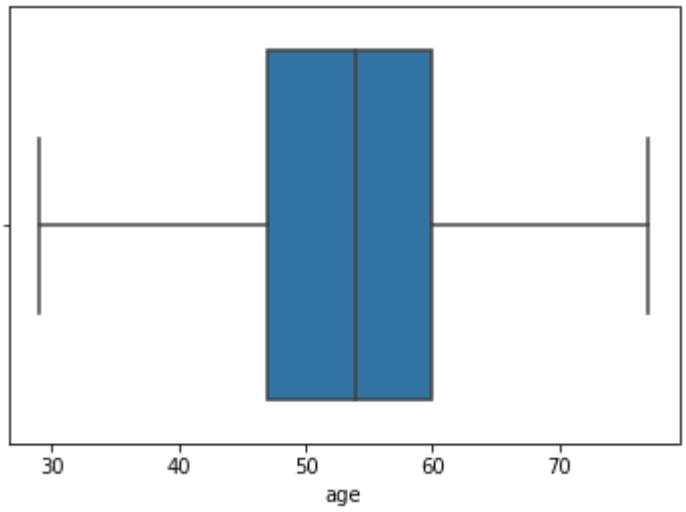
```
# Data model building
```

In [22]:

```
#Detecting Outliers
# 1. Detecting Outliers using IQR (InterQuartile Range)
sns.boxplot(x=df['age'])
#No Outliers observed in 'age'
```

Out[22]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x57fb750>
```

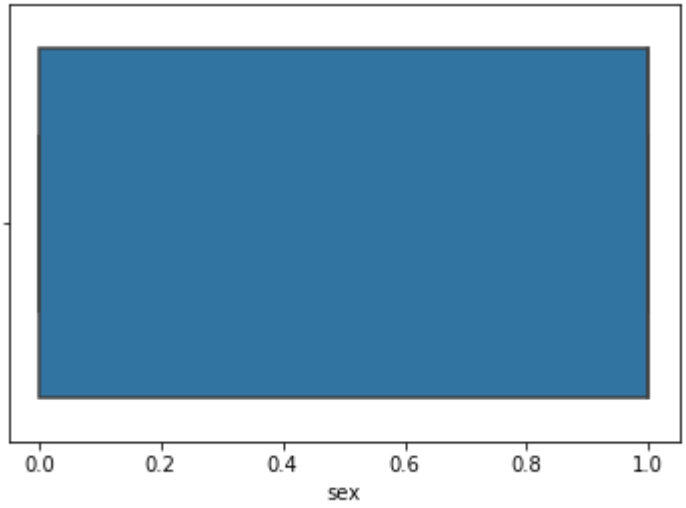


In [24]:

```
sns.boxplot(x=df['sex'])
#No outliers observed in sex data
```

Out[24]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x586f3d0>
```

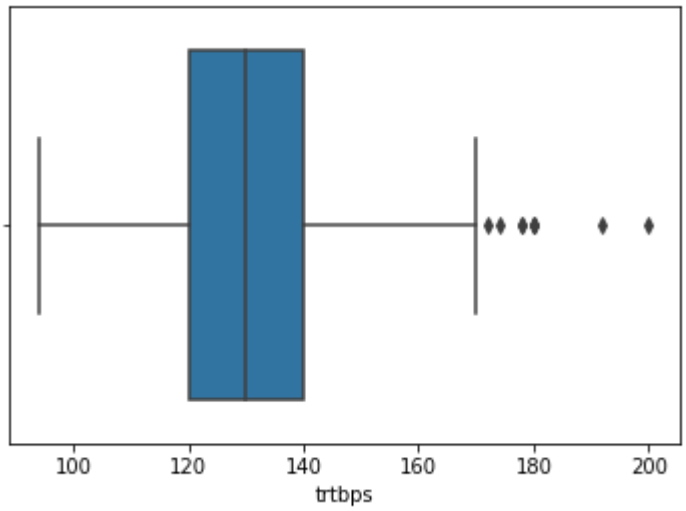


In [25]:

```
sns.boxplot(x=df['trtbps'])
#Some outliers are observed in 'trtbps'. They will be removed later
```

Out[25]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x58aacd0>
```



In [26]:

```
sns.boxplot(x=df['chol'])
#Some outliers are observed in 'chol'. They will be removed later
```

Out[26]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x58e2730>
```