Heart Disease Analysis with EDA using Python

Importing Libraries

In [3]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [4]:
```
import warnings
warnings.filterwarnings('ignore')
```

Import Dataset

In [6]:
```
df = pd.read_csv(r'C:\Users\prane\Downloads\Data Science Nareshit Class notes\Oc
```

In [7]:
```
df
```

Out[7]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	tl
0	63	1	3	145	233	1	0	150	0	2.3	0	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	0	
3	56	1	1	120	236	0	1	178	0	0.8	2	0	
4	57	0	0	120	354	0	1	163	1	0.6	2	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
298	57	0	0	140	241	0	1	123	1	0.2	1	0	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	

303 rows × 14 columns

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

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

Out[9]:
```
Index(['age', 'sex', 'cp', 'trestbps', 'chol', 'fbs', 'restecg', 'thalach',
       'exang', 'oldpeak', 'slope', 'ca', 'thal', 'target'],
      dtype='object')
```

In [10]:
```
df.describe()
```

Out[10]:

	age	sex	cp	trestbps	chol	fbs	reste
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000(
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528(
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525ε
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000(
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000(
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000(
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000(
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000(

In [11]:
```
df.head()
```

Out[11]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2

In [12]:
```
df.tail()
```

Out[12]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	tl
298	57	0	0	140	241	0	1	123	1	0.2	1	0	
299	45	1	3	110	264	0	1	132	0	1.2	1	0	
300	68	1	0	144	193	1	1	141	0	3.4	1	2	
301	57	1	0	130	131	0	1	115	1	1.2	1	1	
302	57	0	1	130	236	0	0	174	0	0.0	1	1	

In [13]:
```
df.isnull()
```

Out[13]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
298	False	False	False	False	False	False	False	False	False	False	False
299	False	False	False	False	False	False	False	False	False	False	False
300	False	False	False	False	False	False	False	False	False	False	False
301	False	False	False	False	False	False	False	False	False	False	False
302	False	False	False	False	False	False	False	False	False	False	False

303 rows × 14 columns

In [14]:
```
df.isnull().isnull()
```

Out[14]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope
0	False	False	False	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...	...	...	...
298	False	False	False	False	False	False	False	False	False	False	False
299	False	False	False	False	False	False	False	False	False	False	False
300	False	False	False	False	False	False	False	False	False	False	False
301	False	False	False	False	False	False	False	False	False	False	False
302	False	False	False	False	False	False	False	False	False	False	False

303 rows × 14 columns

In [15]:
```
df.dtypes
```

Out[15]:
```
age           int64
sex           int64
cp            int64
trestbps      int64
chol          int64
fbs           int64
restecg       int64
thalach       int64
exang         int64
oldpeak     float64
slope         int64
ca            int64
thal          int64
target        int64
dtype: object
```

Univariate Analysis

In [17]:
```
df['target'].nunique()
```

Out[17]:
```
2
```

In [18]:
```
df['target'].unique()
```

Out[18]:
```
array([1, 0], dtype=int64)
```

The unique values are 1 and 0 (1 stands for presence of heart disease and 0 stands for absence of heart disease)

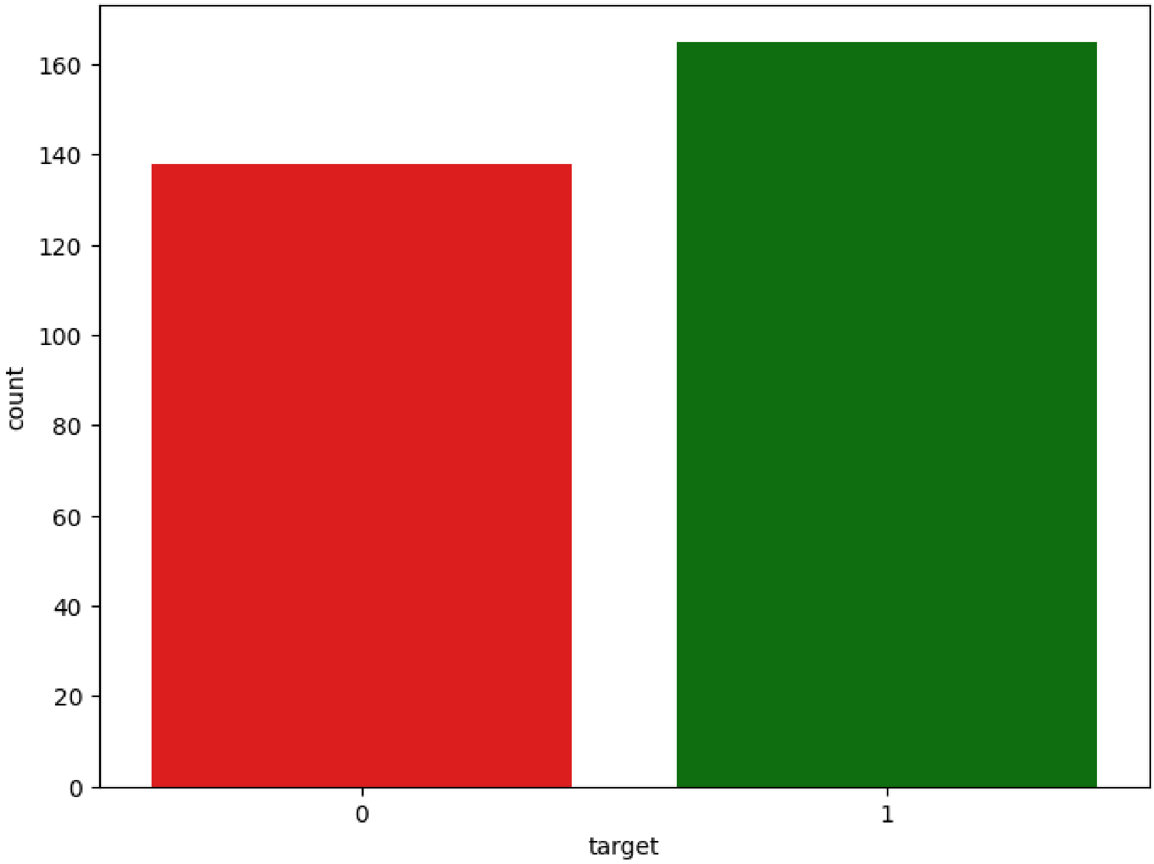
Frequency Distribution of target variable

In [21]:
```
df['target'].value_counts()
```

Out[21]:
```
target
1    165
0    138
Name: count, dtype: int64
```

Visualize Frequency Distribution of `target` variable

In [23]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='target',data=df,palette=['red', 'Green'])
plt.show()
```



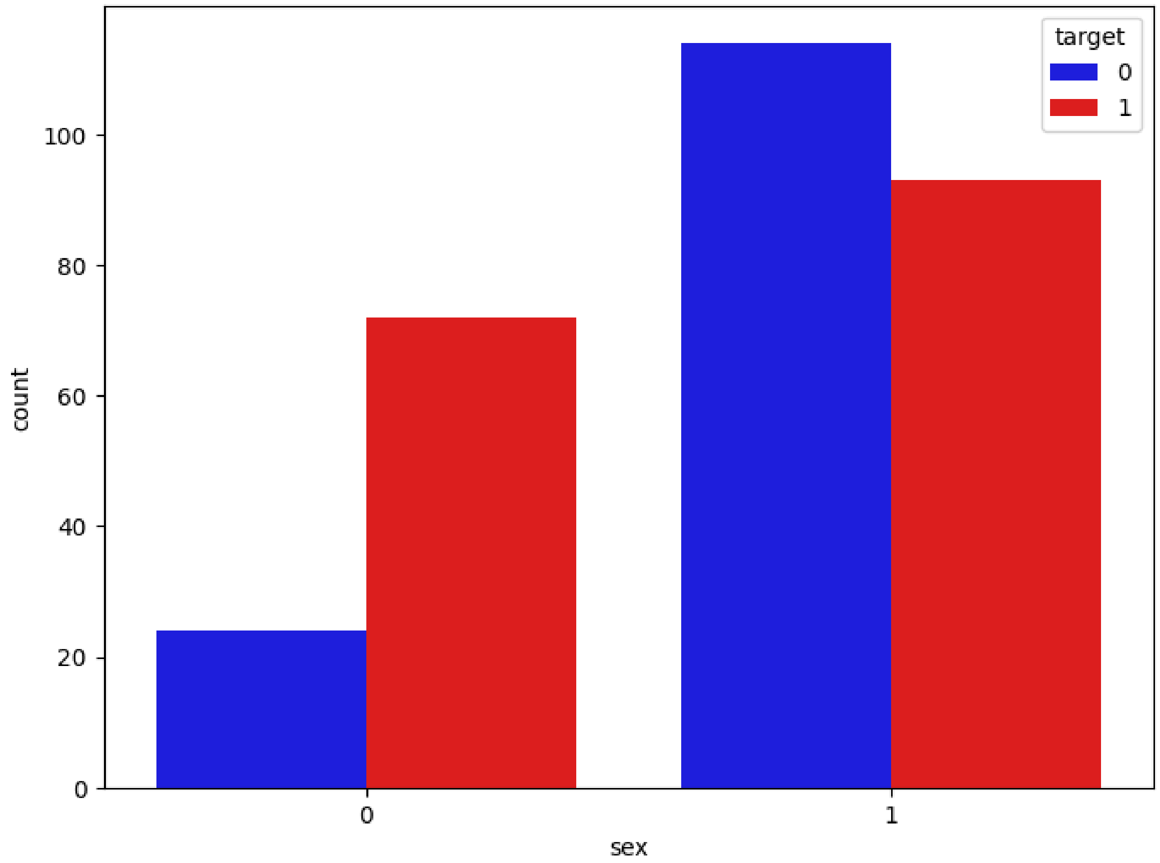
Frequency distribution of `target` variable wrt `sex`

In [25]:
```
df.groupby('sex')['target'].value_counts()
```

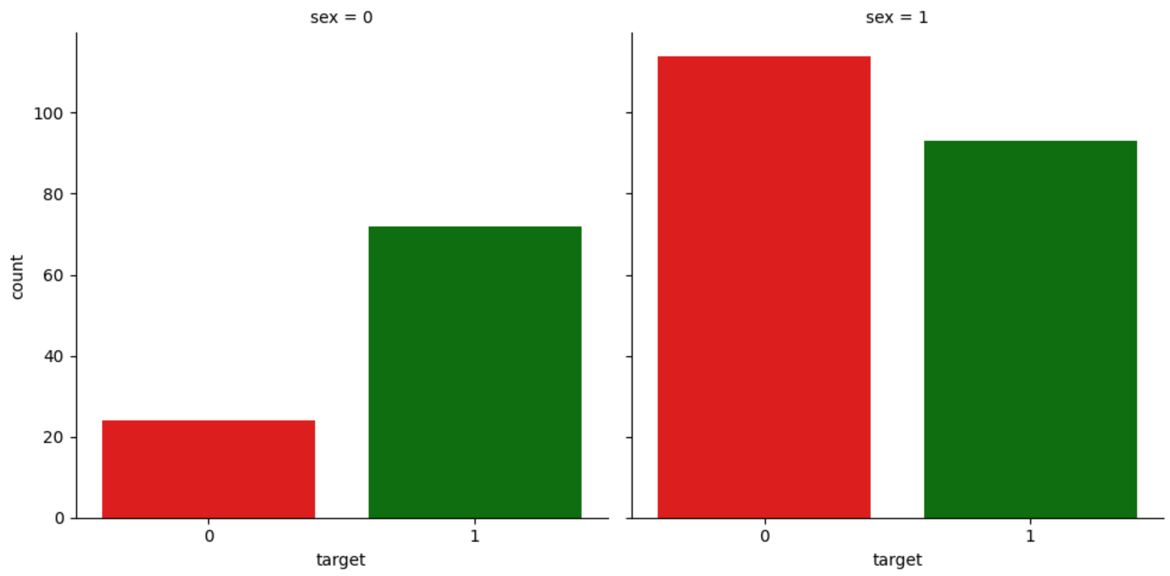
Out[25]:
```
sex  target
0    1          72
     0          24
1    0         114
     1          93
Name: count, dtype: int64
```

sex variable contains 2 integer value 1= male and 2 = female

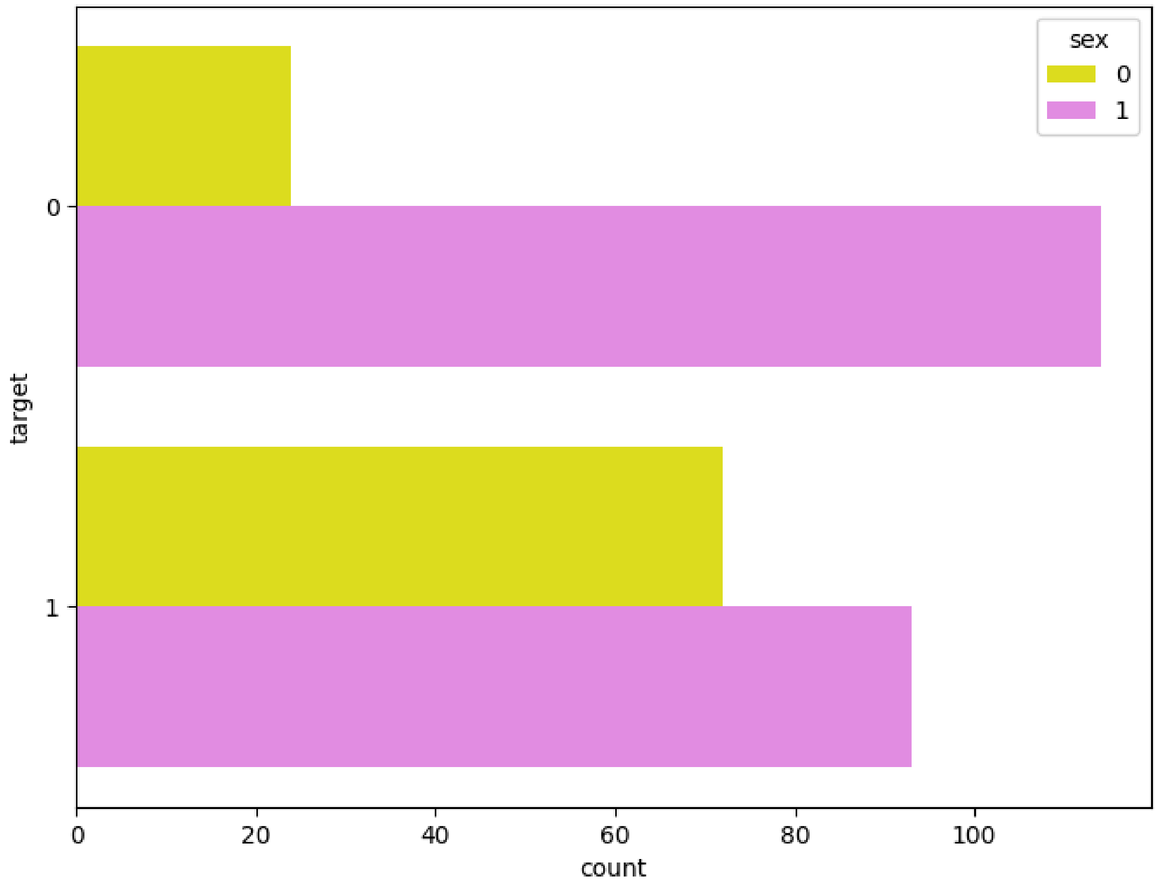
In [27]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='sex',hue='target',data=df,palette=['blue','red'])
plt.show()
```



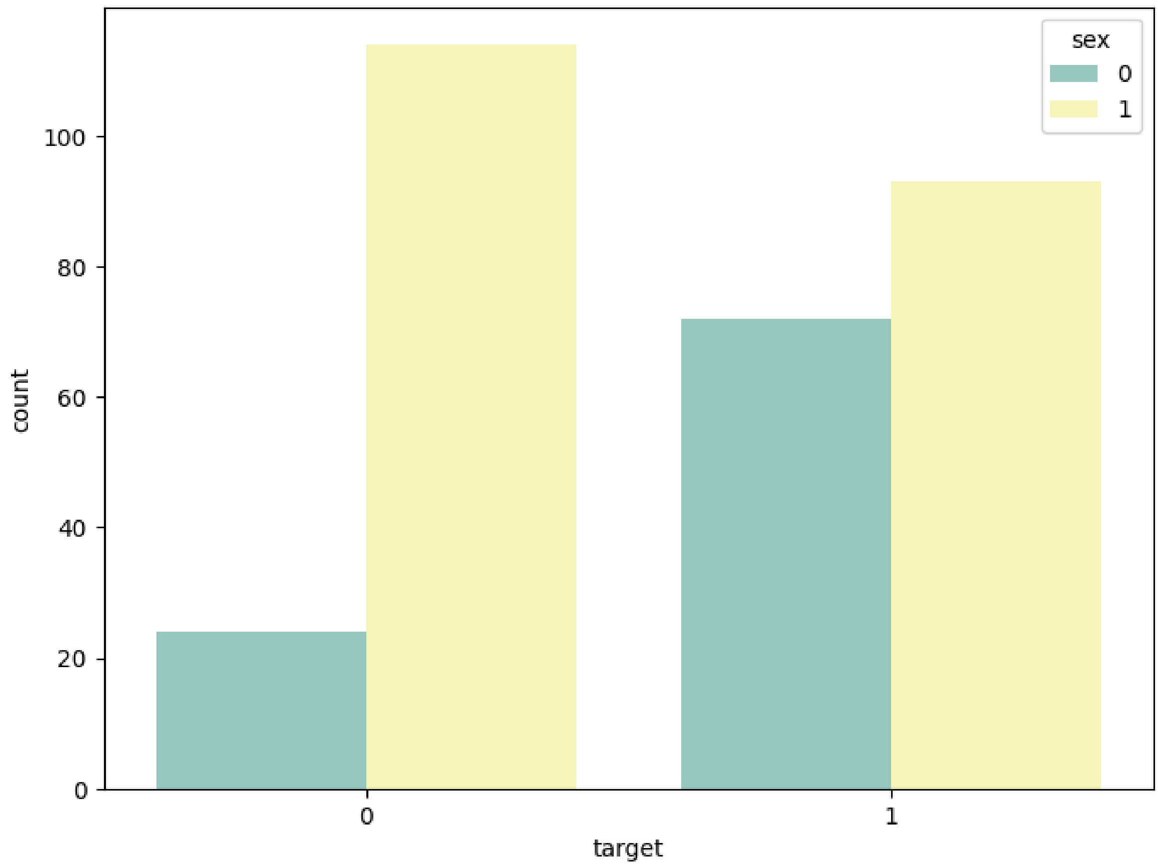
In [28]:
```
ax = sns.catplot(x='target',col='sex',data=df,kind='count',height=5,aspect=1,pal
```



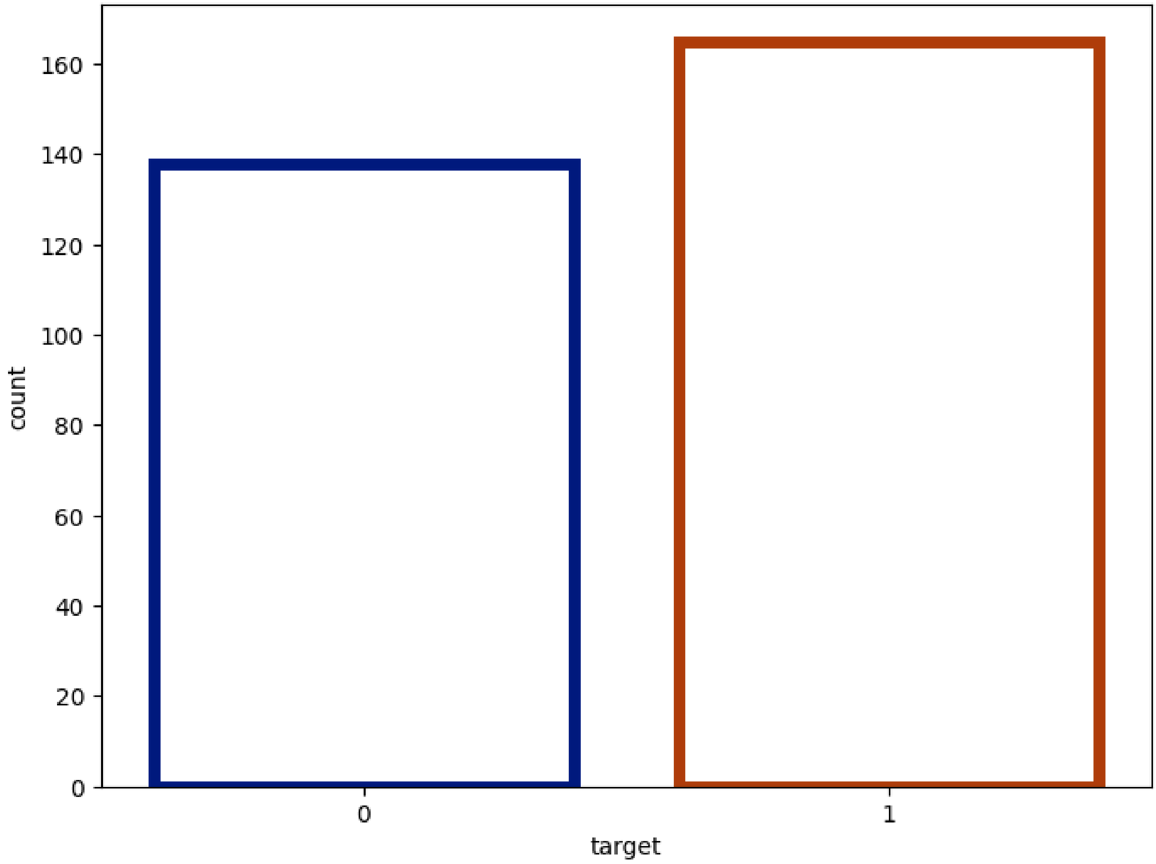
In [29]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(hue='sex',y='target',data=df,palette=['yellow','violet'])
plt.show()
```



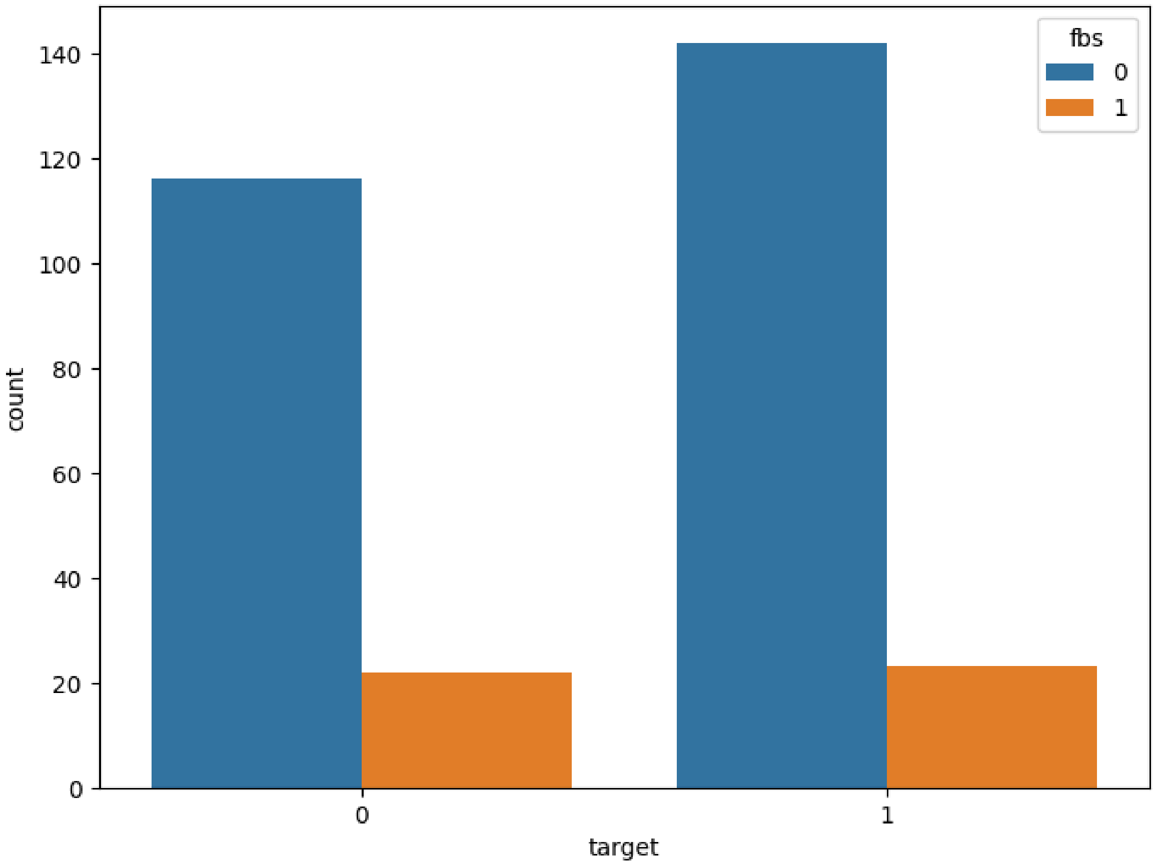
In [30]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='target',hue='sex',data=df,palette='Set3')
plt.show()
```



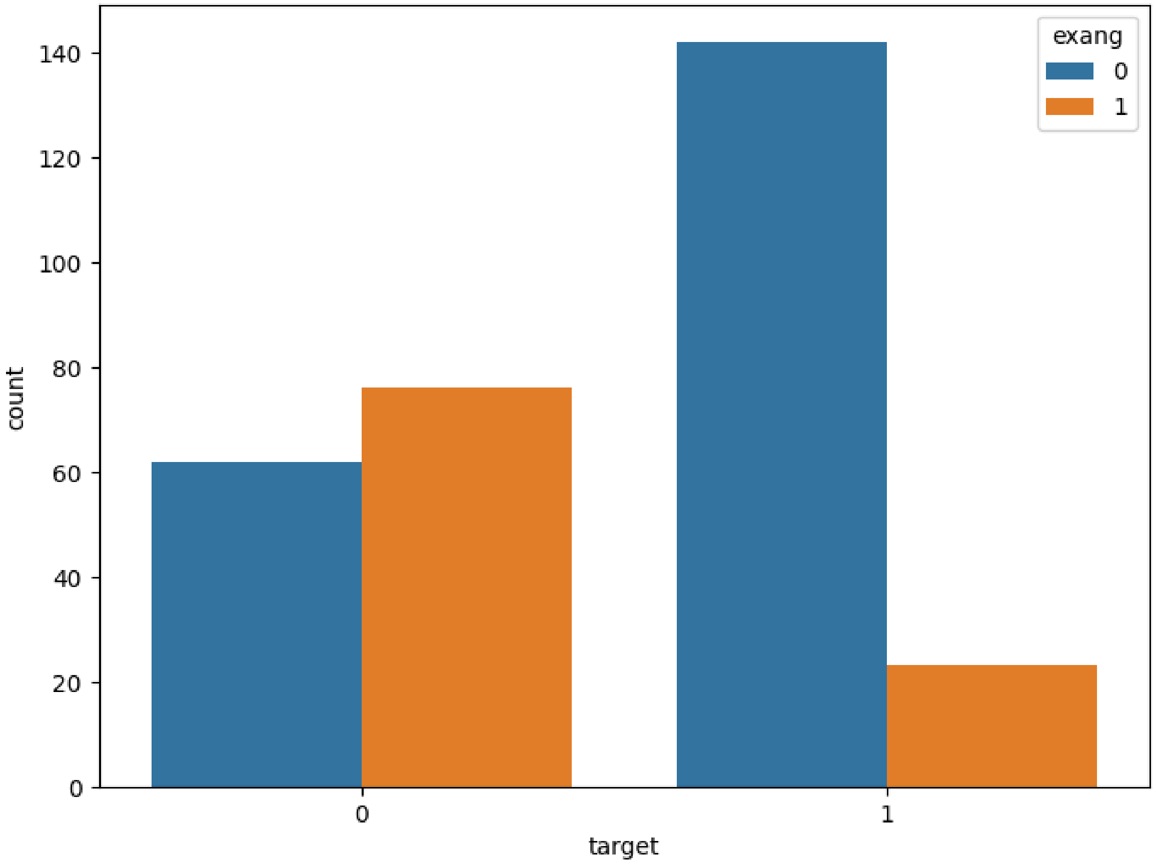
In [31]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='target',data=df,facecolor=(0,0,0,0),linewidth=5, edgecolor
plt.show()
```



In [32]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='target',hue='fbs',data=df)
plt.show()
```



In [33]:
```
f, ax = plt.subplots(figsize=(8,6))
ax = sns.countplot(x='target',hue='exang',data=df)
plt.show()
```



Bivariante Analysis

Estimate correlation coefficients

In [36]:
```
correlation = df.corr()
```

In [37]:
```
correlation['target'].sort_values(ascending=False)
```

Out[37]:
```
target      1.000000
cp          0.433798
thalach     0.421741
slope       0.345877
restecg     0.137230
fbs        -0.028046
chol       -0.085239
trestbps   -0.144931
age        -0.225439
sex        -0.280937
thal       -0.344029
ca         -0.391724
oldpeak    -0.430696
exang      -0.436757
Name: target, dtype: float64
```

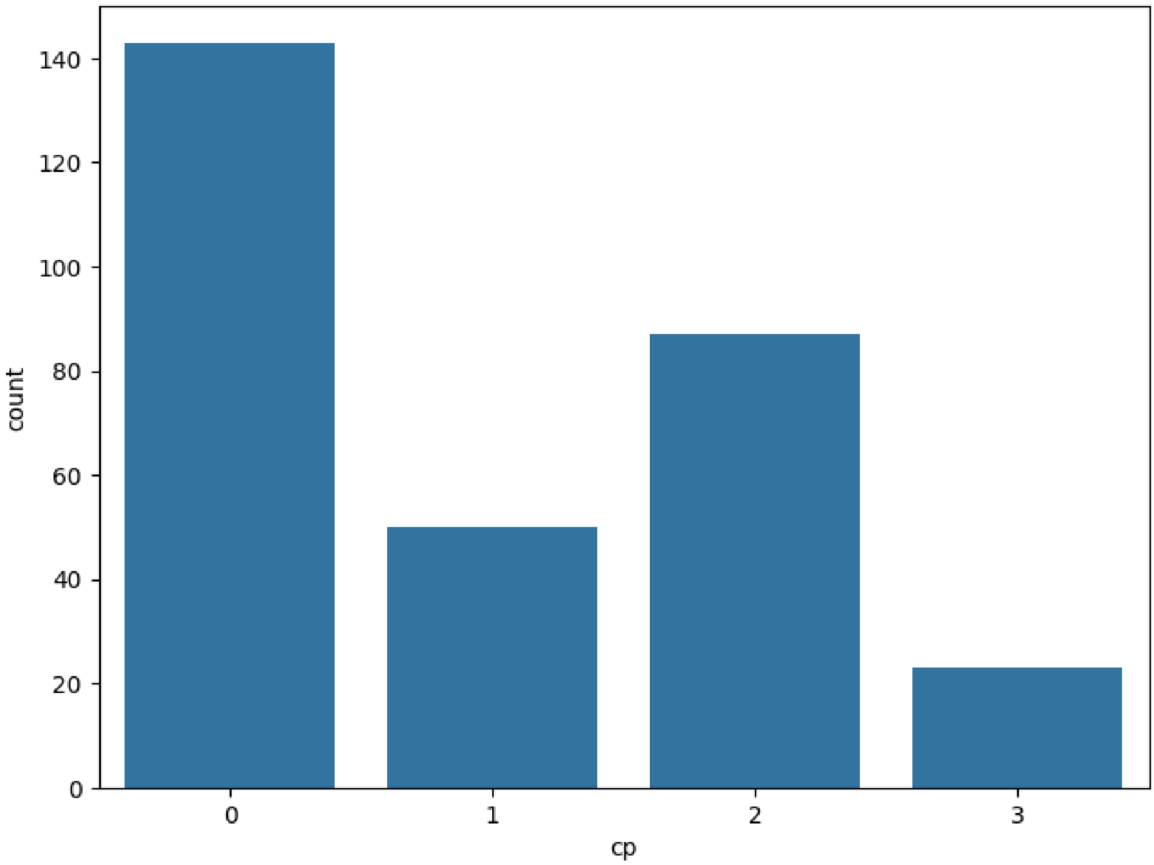
In [66]:
```
df['cp'].nunique()
```

Out[66]:
```
4
```

In [68]:
```
df['cp'].value_counts()
```

Out[68]:
```
cp
0    143
2     87
1     50
3     23
Name: count, dtype: int64
```

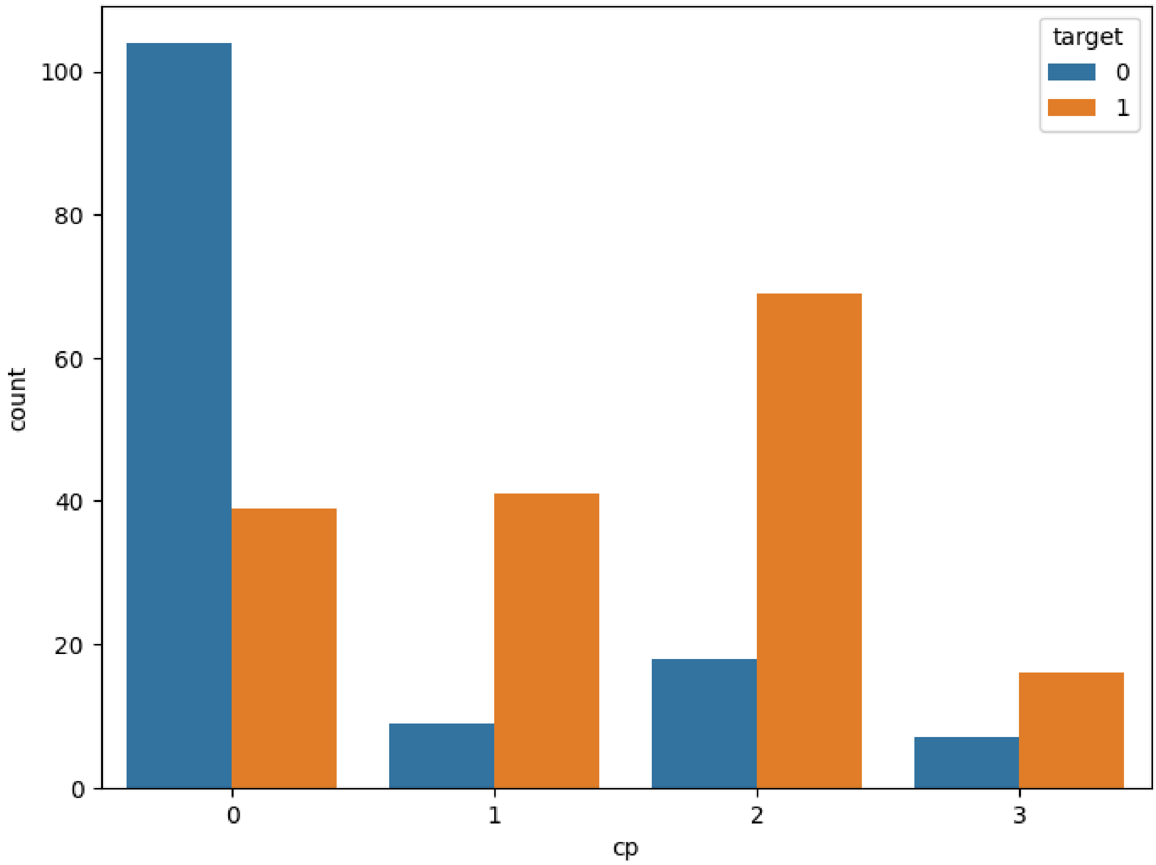
In [70]:
```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.countplot(x="cp", data=df)
plt.show()
```



In [72]:
```
df.groupby('cp')['target'].value_counts()
```

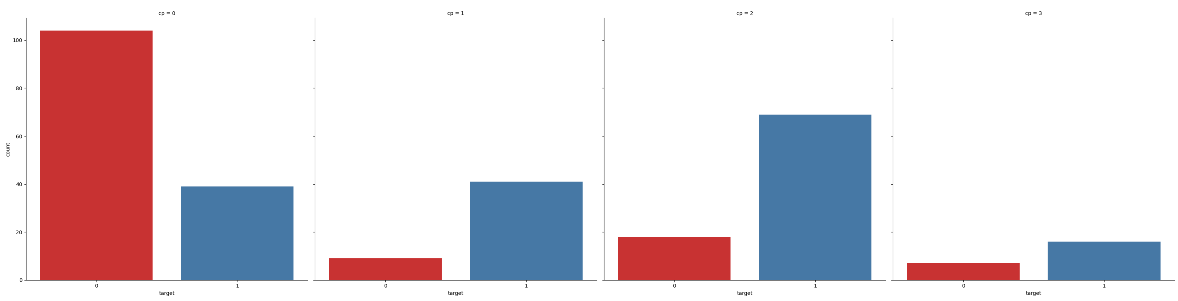
Out[72]:
```
cp  target
0   0         104
    1          39
1   1          41
    0           9
2   1          69
    0          18
3   1          16
    0           7
Name: count, dtype: int64
```

In [74]:
```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.countplot(x="cp", hue="target", data=df)
plt.show()
```



In [102…

```
ax = sns.catplot(x="target", col="cp", data=df, kind="count", height=8, aspect=1
```



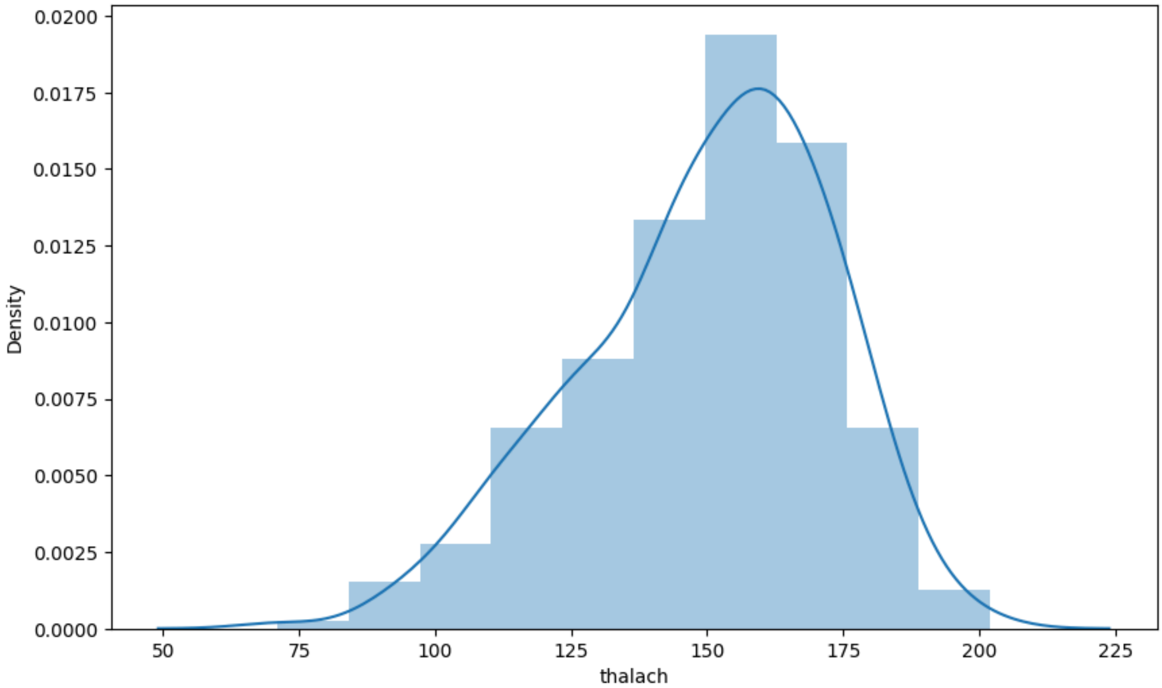
In [78]:

```
df['thalach'].nunique()
```

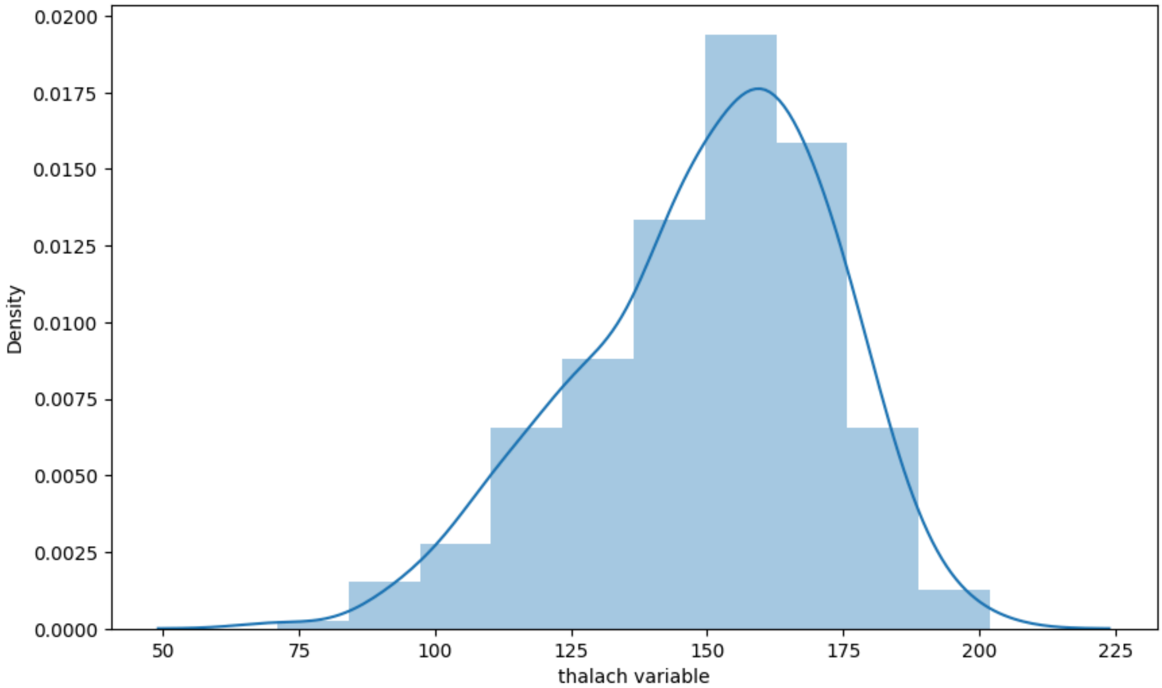
Out[78]: 91

In [80]:

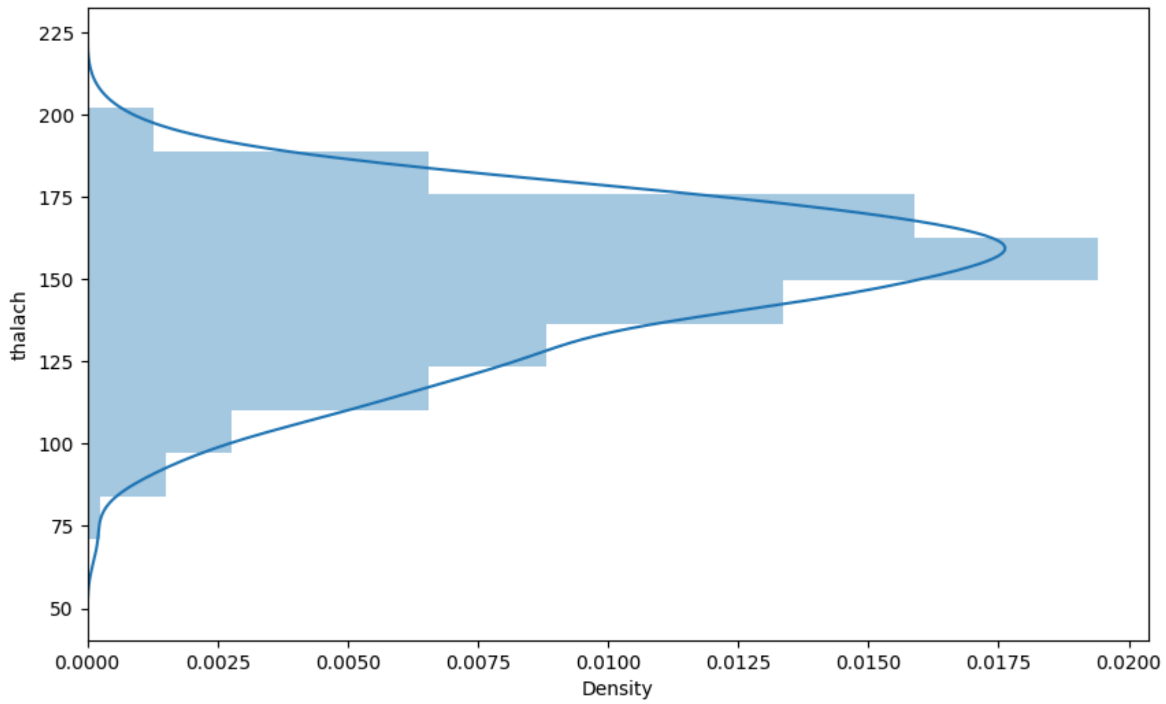
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
ax = sns.distplot(x, bins=10)
plt.show()
```



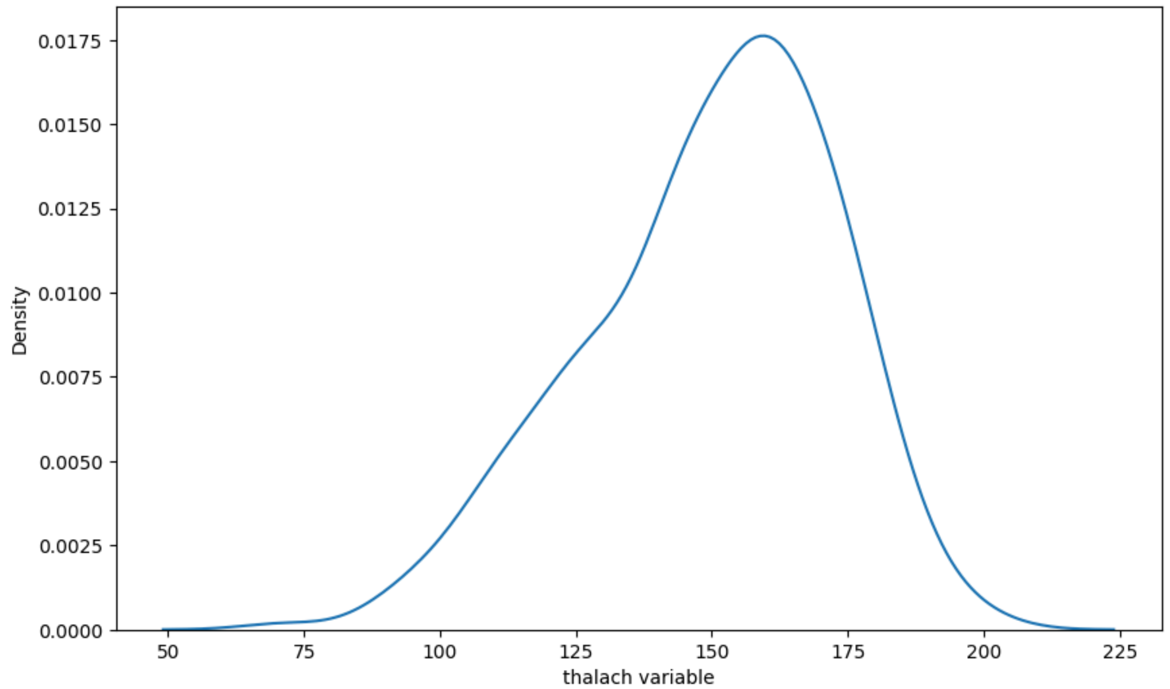
In [82]:
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
x = pd.Series(x, name="thalach variable")
ax = sns.distplot(x, bins=10)
plt.show()
```



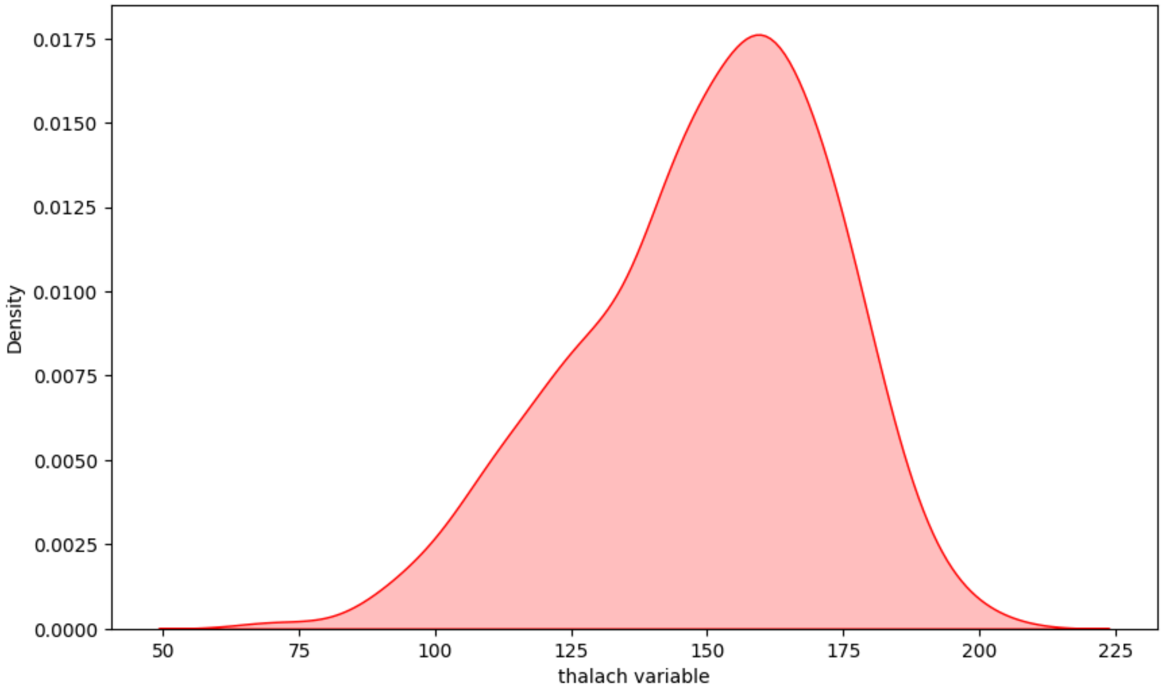
In [84]:
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
ax = sns.distplot(x, bins=10, vertical=True)
plt.show()
```



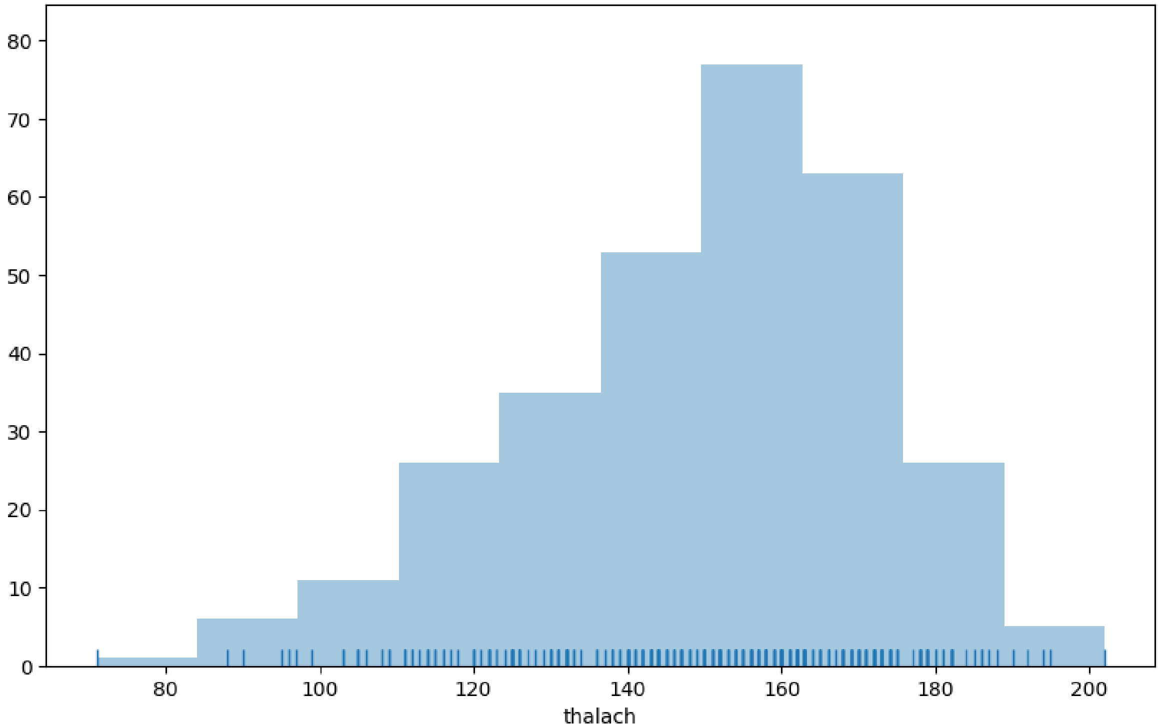
In [86]:
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
x = pd.Series(x, name="thalach variable")
ax = sns.kdeplot(x)
plt.show()
```



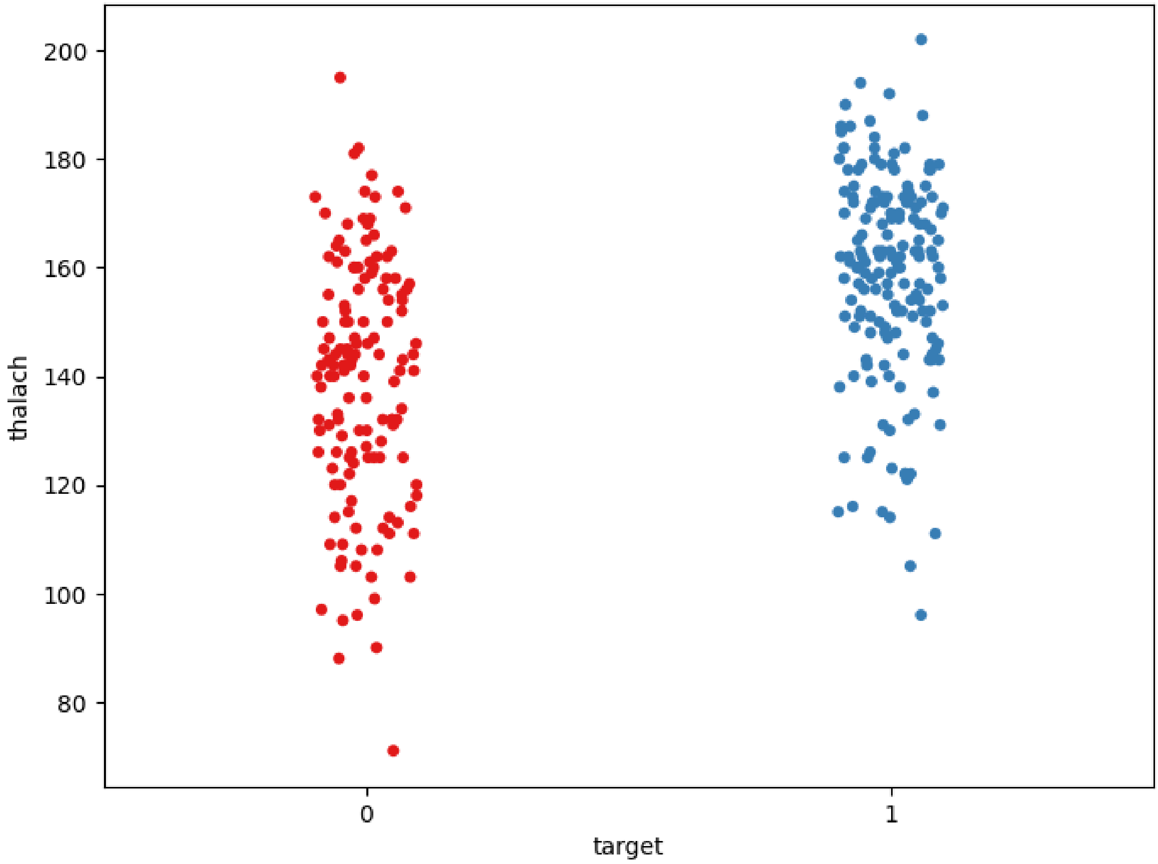
In [88]:
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
x = pd.Series(x, name="thalach variable")
ax = sns.kdeplot(x, shade=True, color='r')
plt.show()
```



In [90]:
```
f, ax = plt.subplots(figsize=(10,6))
x = df['thalach']
ax = sns.distplot(x, kde=False, rug=True, bins=10)
plt.show()
```

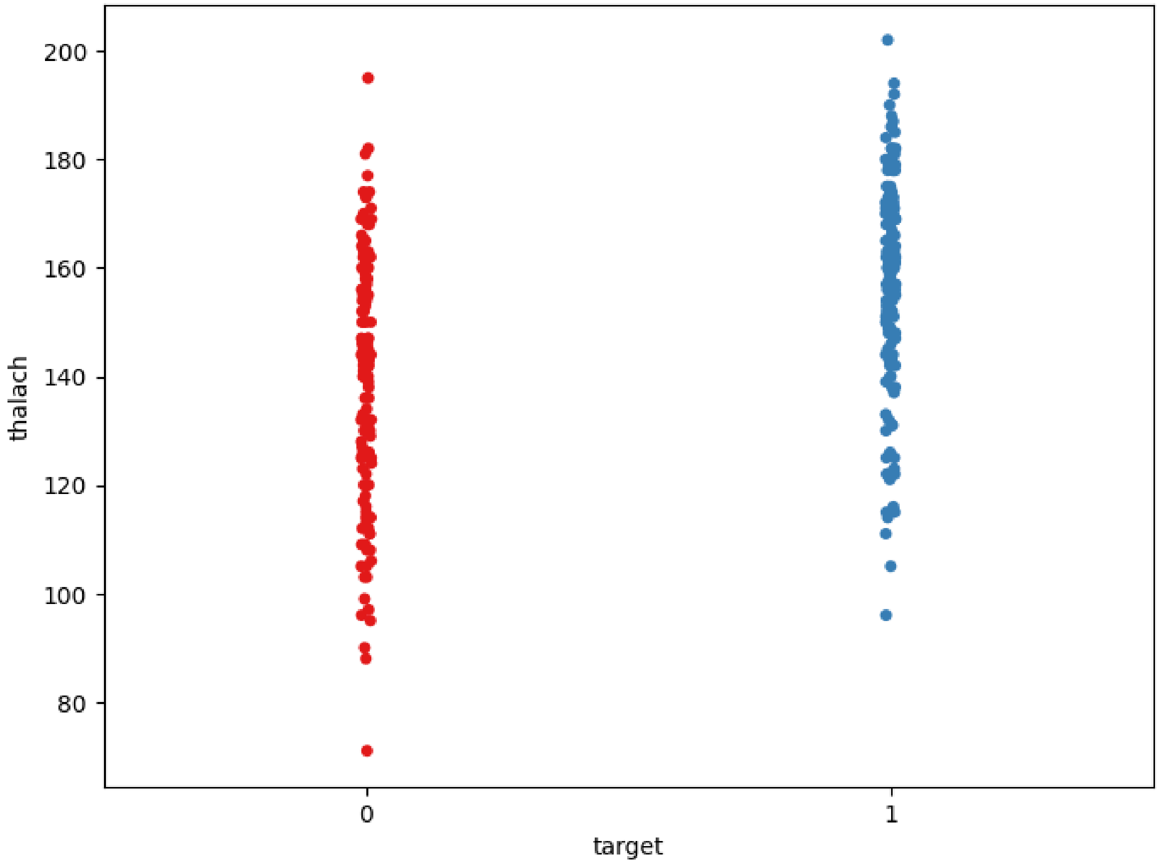


In [100…
```
f, ax = plt.subplots(figsize=(8, 6))
sns.stripplot(x="target", y="thalach", data=df,palette="Set1")
plt.show()
```



In [106…

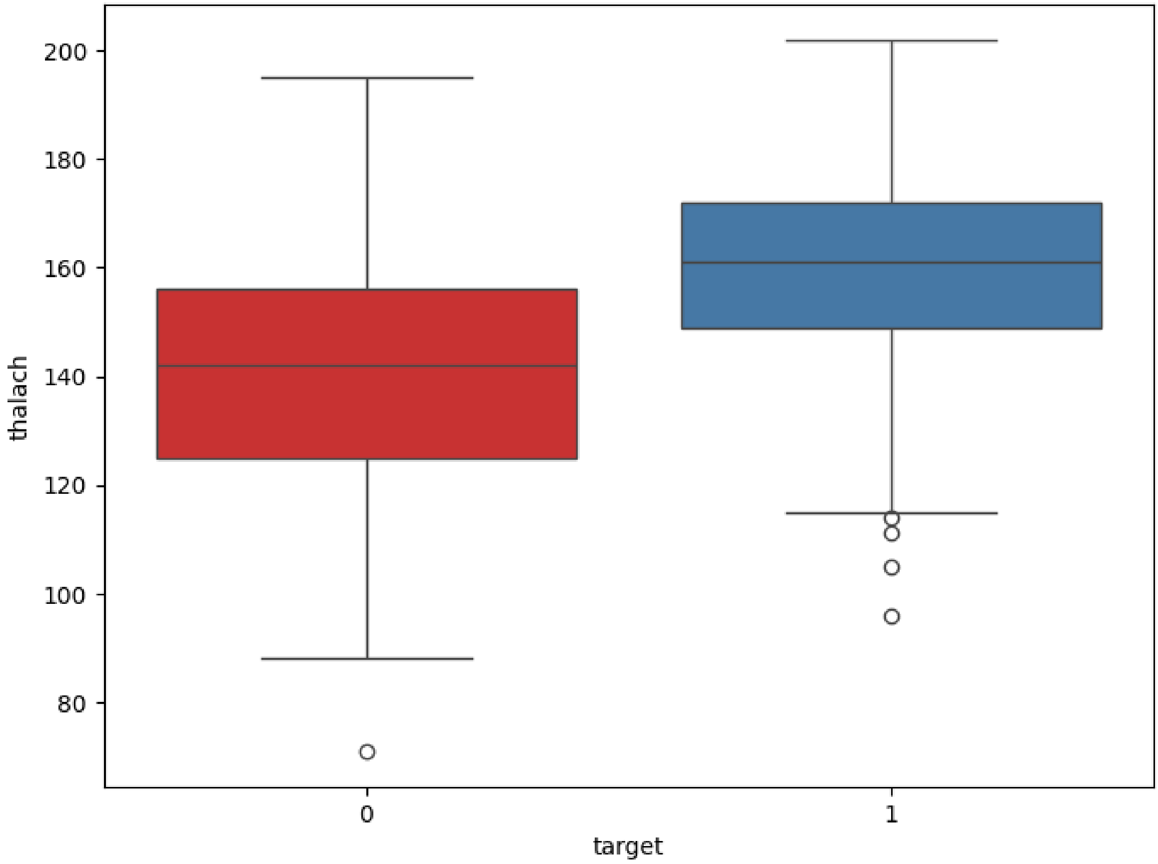
```
f, ax = plt.subplots(figsize=(8, 6))
sns.stripplot(x="target", y="thalach", data=df, jitter = 0.01,palette='Set1')
plt.show()
```



In [110…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x="target", y="thalach", data=df, palette='Set1')
```

```
plt.show()
```



Multivariate analysis

- The objective of the multivariate analysis is to discover patterns and relationships in the dataset.

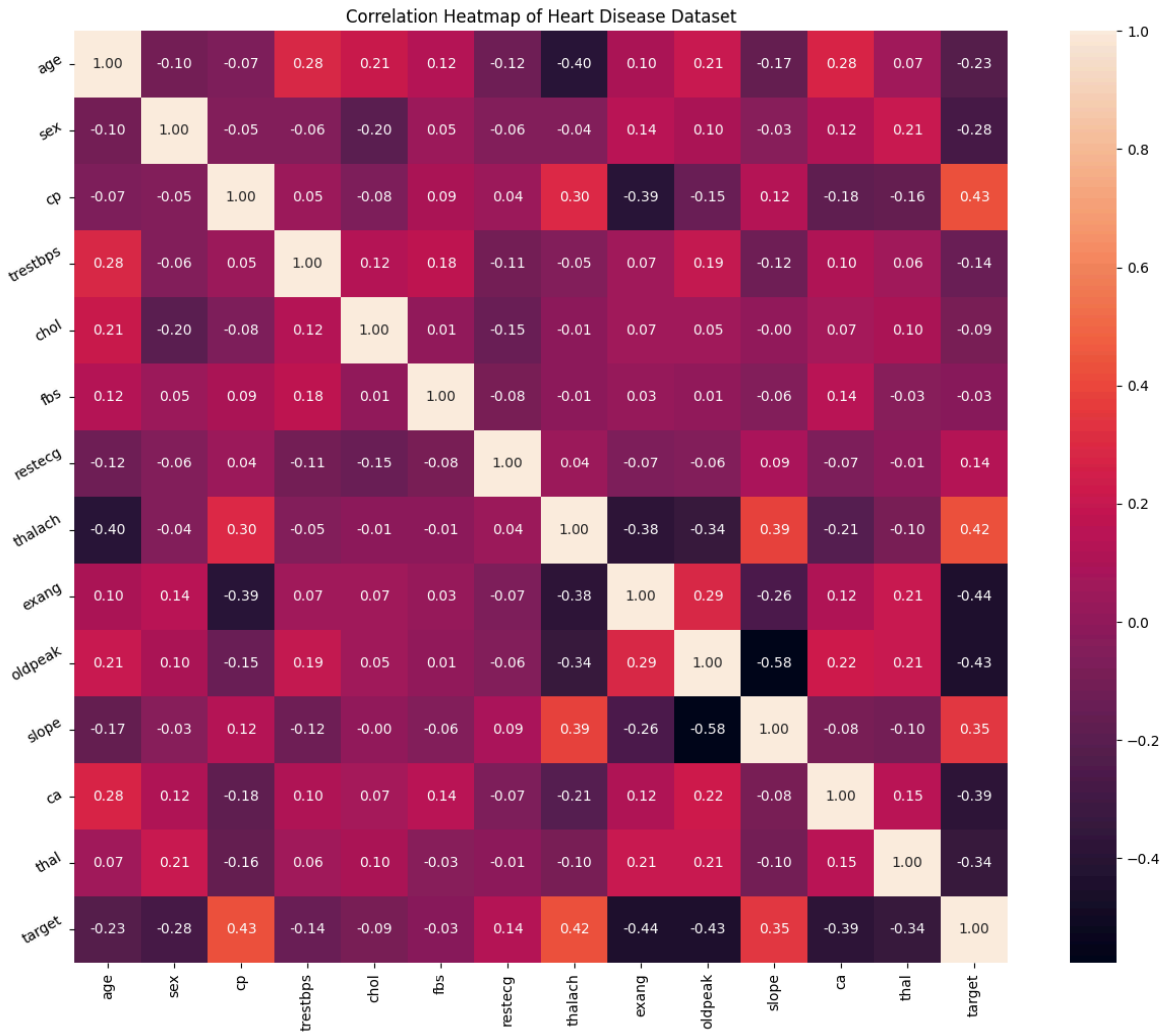
Discover patterns and relationships

- An important step in EDA is to discover patterns and relationships between variables in the dataset.
- I will use `heat map` and `pair plot` to discover the patterns and relationships in the dataset.
- First of all, I will draw a `heat map`.

Heat Map

In [117…

```
plt.figure(figsize=(16,12))
plt.title('Correlation Heatmap of Heart Disease Dataset')
a = sns.heatmap(correlation, square=True, annot=True, fmt='.2f', linecolor='whit
a.set_xticklabels(a.get_xticklabels(), rotation=90)
a.set_yticklabels(a.get_yticklabels(), rotation=30)
plt.show()
```


Correlation Heatmap of Heart Disease Dataset
age 1.00 -0.10 -0.07 0.28 0.21 0.12 -0.12 -0.40 0.10 0.21 -0.17 0.28 0.07 -0.23
sex -0.10 1.00 -0.05 -0.06 -0.20 0.05 -0.06 -0.04 0.14 0.10 -0.03 0.12 0.21 -0.28
cp -0.07 -0.05 1.00 0.05 -0.08 0.09 0.04 0.30 -0.39 -0.15 0.12 -0.18 -0.16 0.43
trestbps 0.28 -0.06 0.05 1.00 0.12 0.18 -0.11 -0.05 0.07 0.19 -0.12 0.10 0.06 -0.14
chol 0.21 -0.20 -0.08 0.12 1.00 0.01 -0.15 -0.01 0.07 0.05 -0.00 0.07 0.10 -0.09
fbs 0.12 0.05 0.09 0.18 0.01 1.00 -0.08 -0.01 0.03 0.01 -0.06 0.14 -0.03 -0.03
restecg -0.12 -0.06 0.04 -0.11 -0.15 -0.08 1.00 0.04 -0.07 -0.06 0.09 -0.07 -0.01 0.14
thalach -0.40 -0.04 0.30 -0.05 -0.01 -0.01 0.04 1.00 -0.38 -0.34 0.39 -0.21 -0.10 0.42
exang 0.10 0.14 -0.39 0.07 0.07 0.03 -0.07 -0.38 1.00 0.29 -0.26 0.12 0.21 -0.44
oldpeak 0.21 0.10 -0.15 0.19 0.05 0.01 -0.06 -0.34 0.29 1.00 -0.58 0.22 0.21 -0.43
slope -0.17 -0.03 0.12 -0.12 -0.00 -0.06 0.09 0.39 -0.26 -0.58 1.00 -0.08 -0.10 0.35
ca 0.28 0.12 -0.18 0.10 0.07 0.14 -0.07 -0.21 0.12 0.22 -0.08 1.00 0.15 -0.39
thal 0.07 0.21 -0.16 0.06 0.10 -0.03 -0.01 -0.10 0.21 0.21 -0.10 0.15 1.00 -0.34
target -0.23 -0.28 0.43 -0.14 -0.09 -0.03 0.14 0.42 -0.44 -0.43 0.35 -0.39 -0.34 1.00
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Interpretation

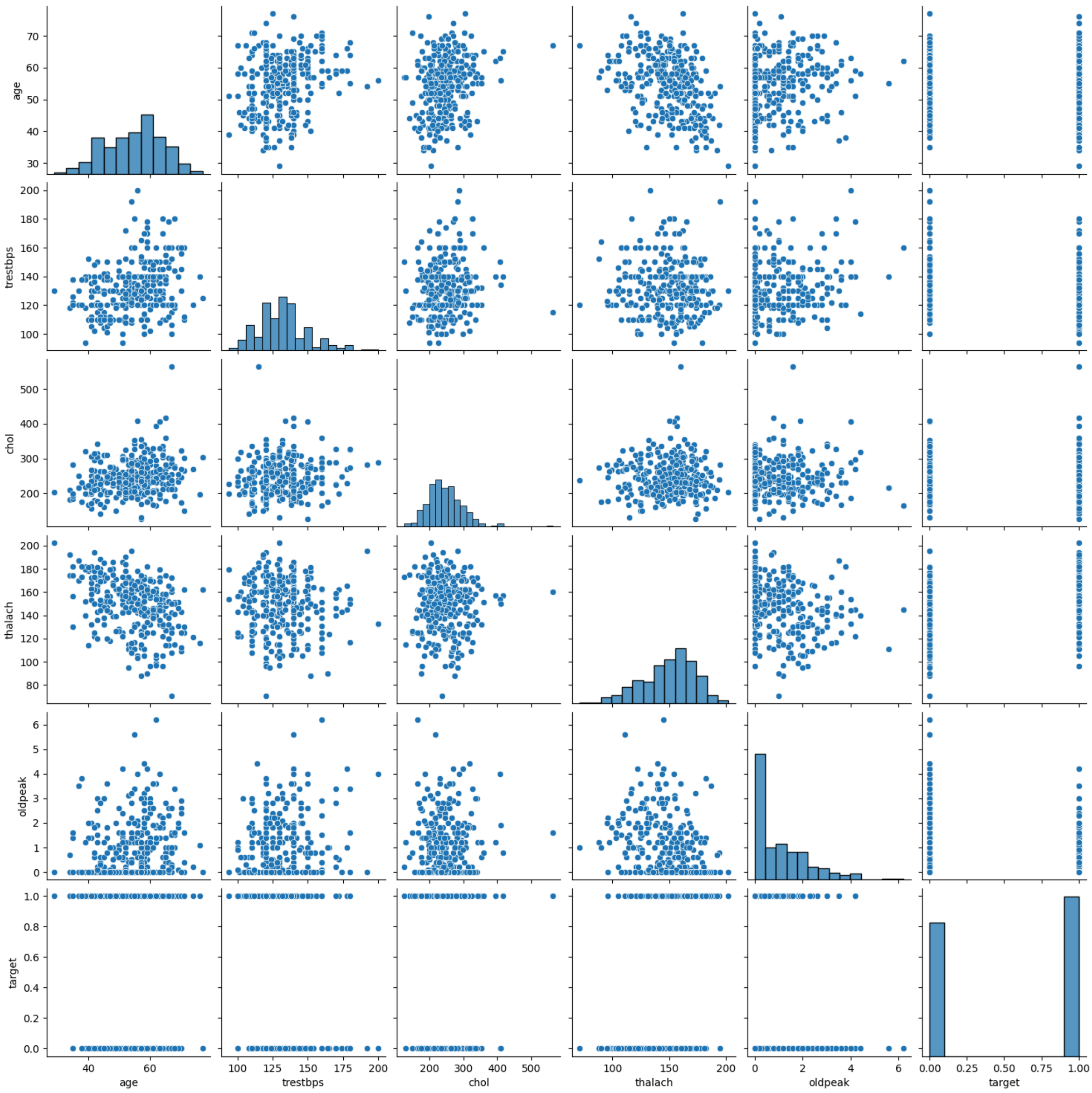
From the above correlation heat map, we can conclude that :-

- `target` and `cp` variable are mildly positively correlated (correlation coefficient = 0.43).
- `target` and `thalach` variable are also mildly positively correlated (correlation coefficient = 0.42).
- `target` and `slope` variable are weakly positively correlated (correlation coefficient = 0.35).
- `target` and `exang` variable are mildly negatively correlated (correlation coefficient = -0.44).
- `target` and `oldpeak` variable are also mildly negatively correlated (correlation coefficient = -0.43).
- `target` and `ca` variable are weakly negatively correlated (correlation coefficient = -0.39).
- `target` and `thal` variable are also waekly negatively correlated (correlation coefficient = -0.34).

Pair Plot

In [125…

```
num_var = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak', 'target' ]
sns.pairplot(df[num_var], kind='scatter', diag_kind='hist')
plt.show()
```



Analysis of `age` and other variables

Check the number of unique values in `age` variable

In [133…

```
df['age'].nunique()
```

Out[133…

```
41
```

View statistical summary of `age` variable

In [136…

```
df['age'].describe()
```

Out[136…

```
count    303.000000
mean      54.366337
std        9.082101
min       29.000000
25%       47.500000
50%       55.000000
75%       61.000000
max       77.000000
Name: age, dtype: float64
```

Interpretation

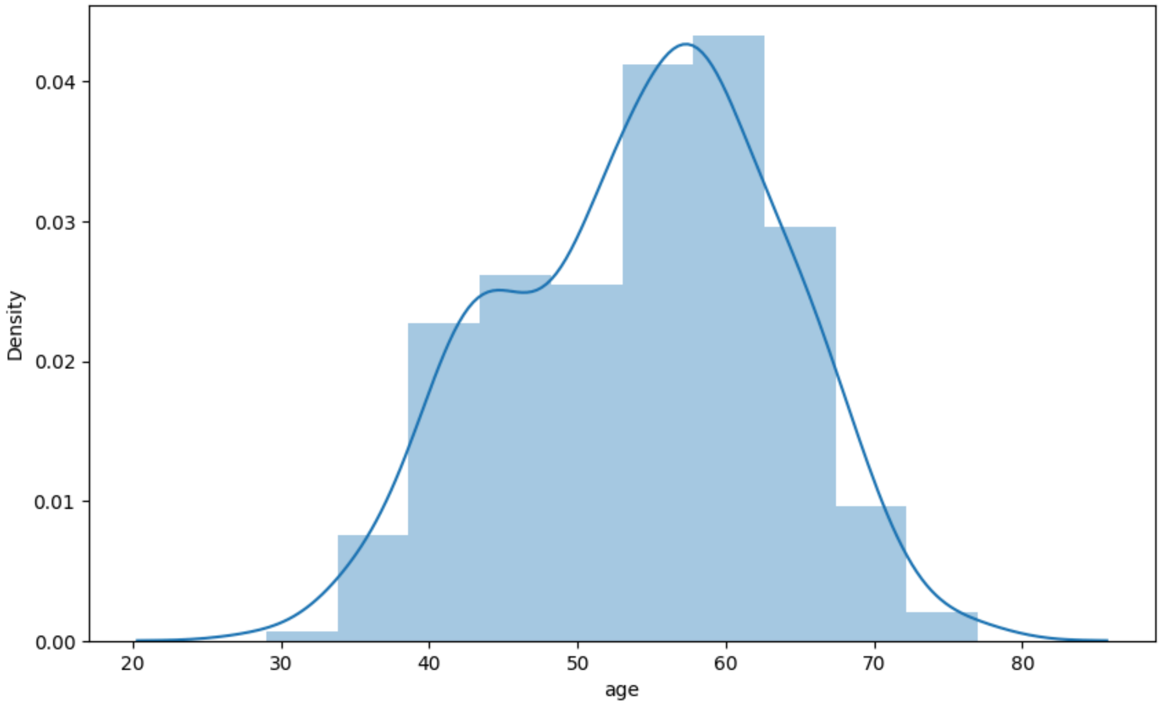
- The mean value of the `age` variable is 54.37 years.
- The minimum and maximum values of `age` are 29 and 77 years.

Plot the distribution of `age` variable

Now, I will plot the distribution of `age` variable to view the statistical properties.

In [142...

```
f, ax = plt.subplots(figsize=(10,6))
x = df['age']
ax = sns.distplot(x, bins=10)
plt.show()
```



Interpretation

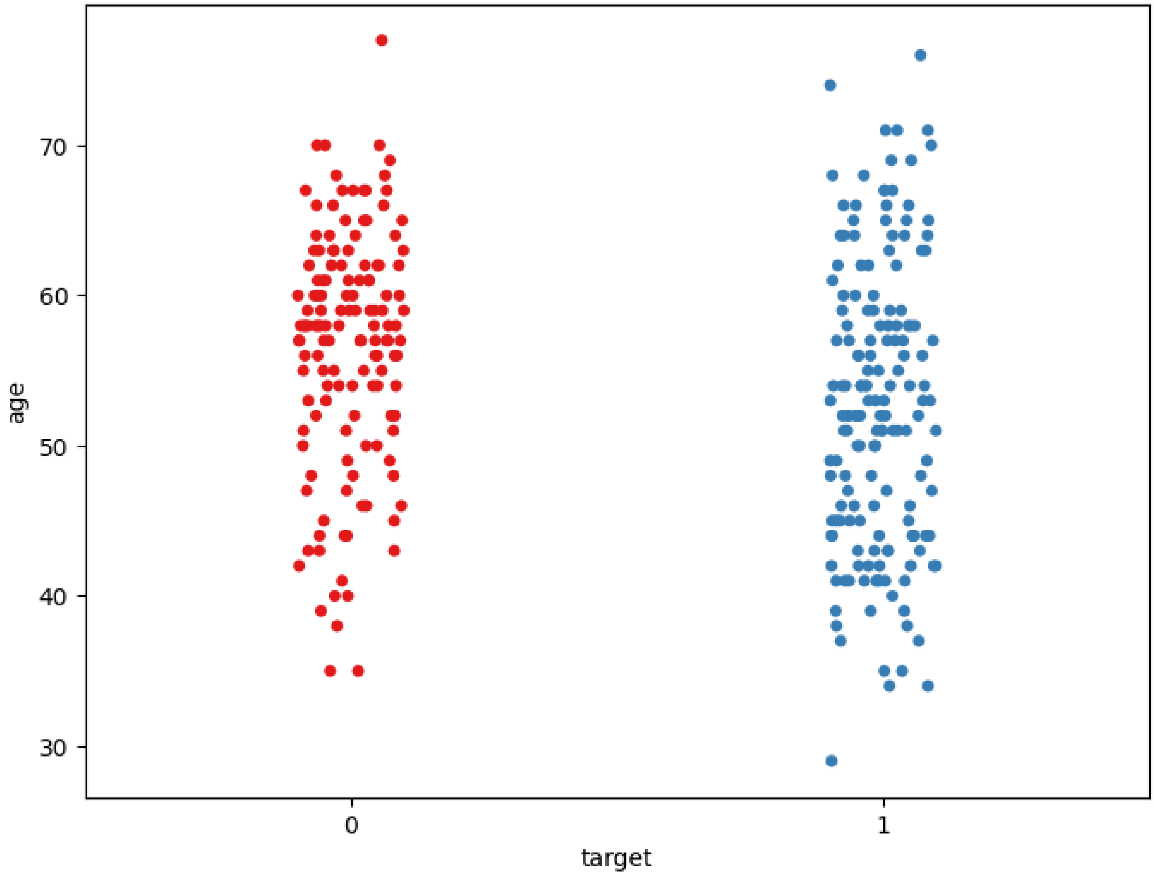
- The `age` variable distribution is approximately normal.

Analyze `age` and `target` variable

Visualize frequency distribution of `age` variable wrt `target`

In [153...

```
f, ax = plt.subplots(figsize=(8, 6))
sns.stripplot(x="target", y="age", data=df, palette='Set1')
plt.show()
```



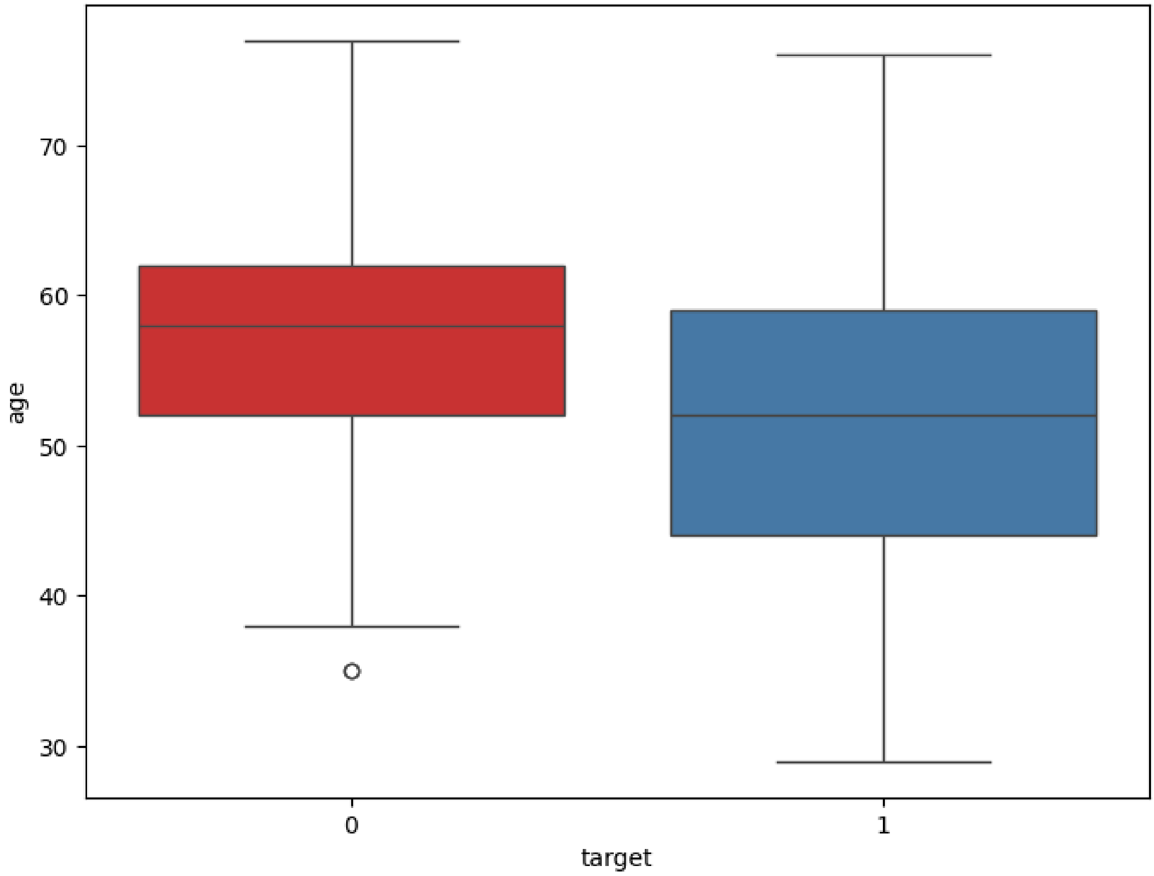
Interpretation

- We can see that the people suffering from heart disease (target = 1) and people who are not suffering from heart disease (target = 0) have comparable ages.

Visualize distribution of `age` variable wrt `target` with boxplot

In [160…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x="target", y="age", data=df, palette='Set1')
plt.show()
```



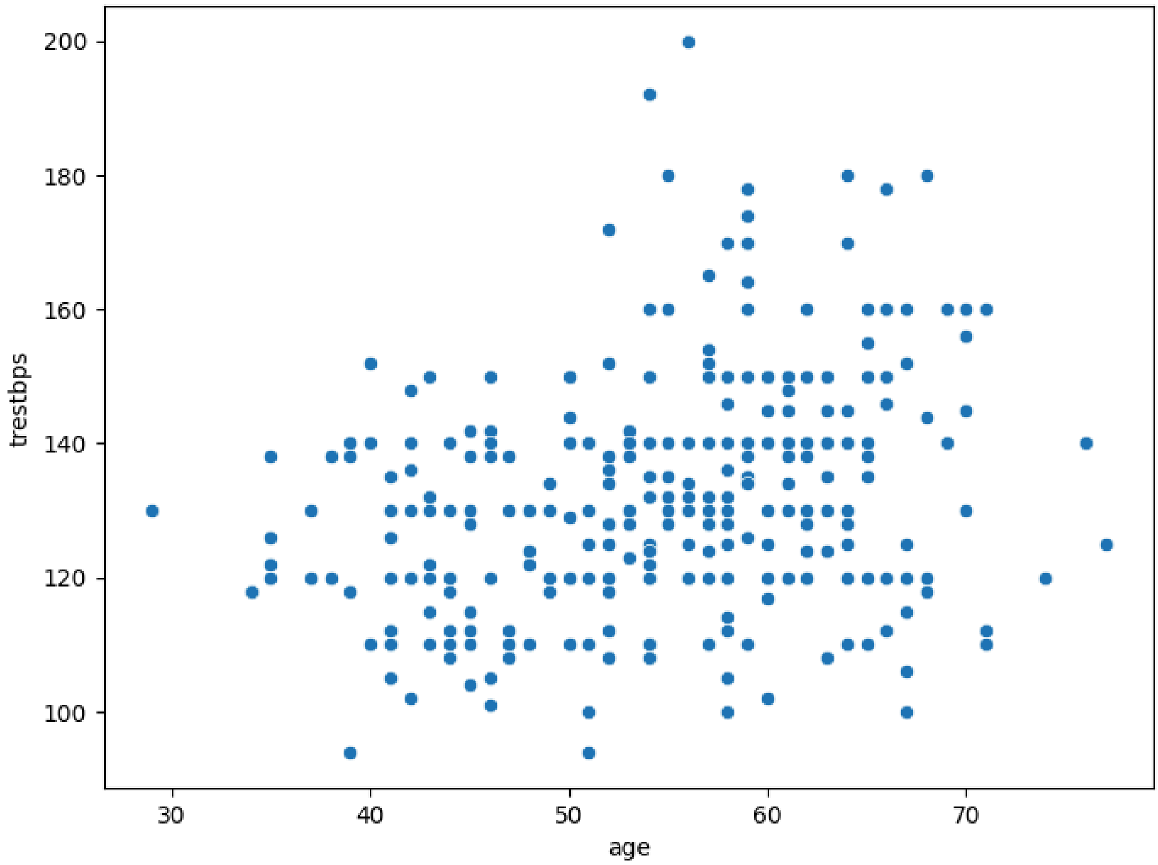
Interpretation

- The above boxplot tells two different things :
  - The mean age of the people who have heart disease is less than the mean age of the people who do not have heart disease.
  - The dispersion or spread of age of the people who have heart disease is greater than the dispersion or spread of age of the people who do not have heart disease.

Analyze `age` and `trestbps` variable

In [166…

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.scatterplot(x="age", y="trestbps", data=df)
plt.show()
```

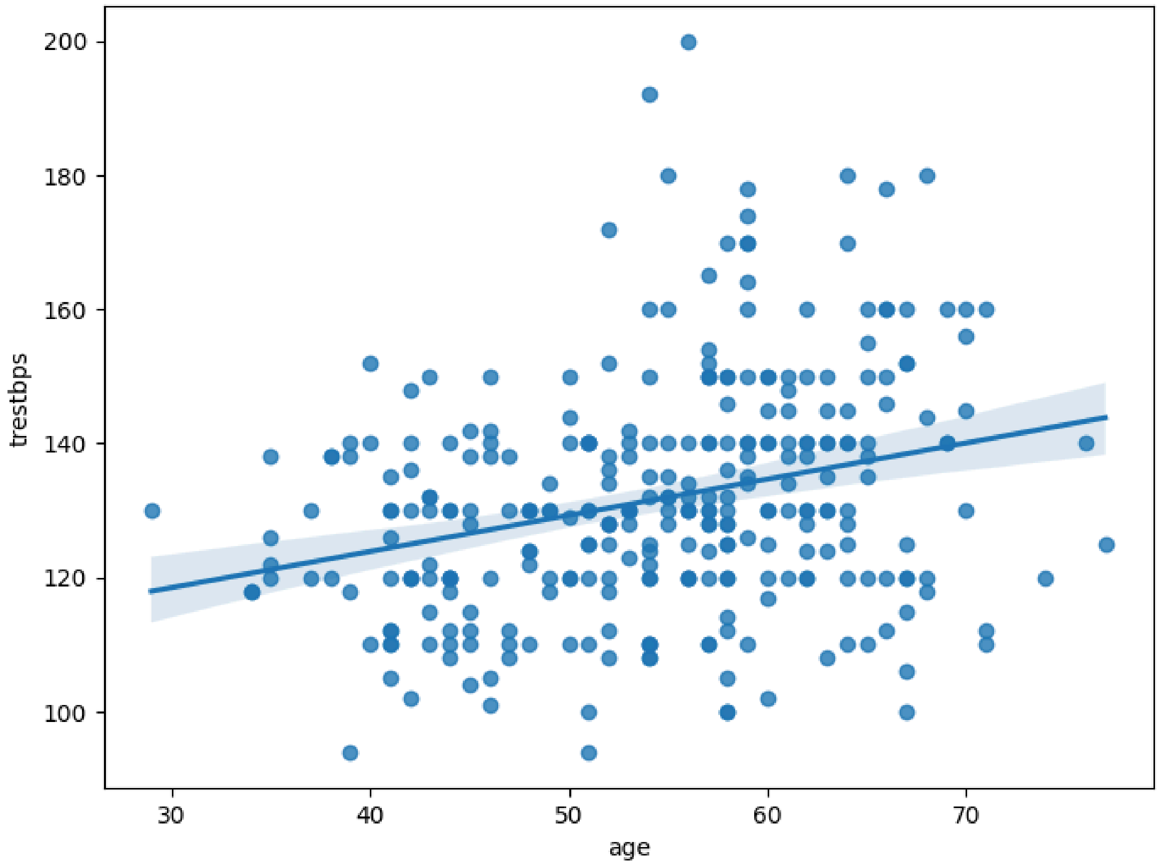


Interpretation

- The above scatter plot shows that there is no correlation between `age` and `trestbps` variable.

In [169...

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.regplot(x="age", y="trestbps", data=df)
plt.show()
```



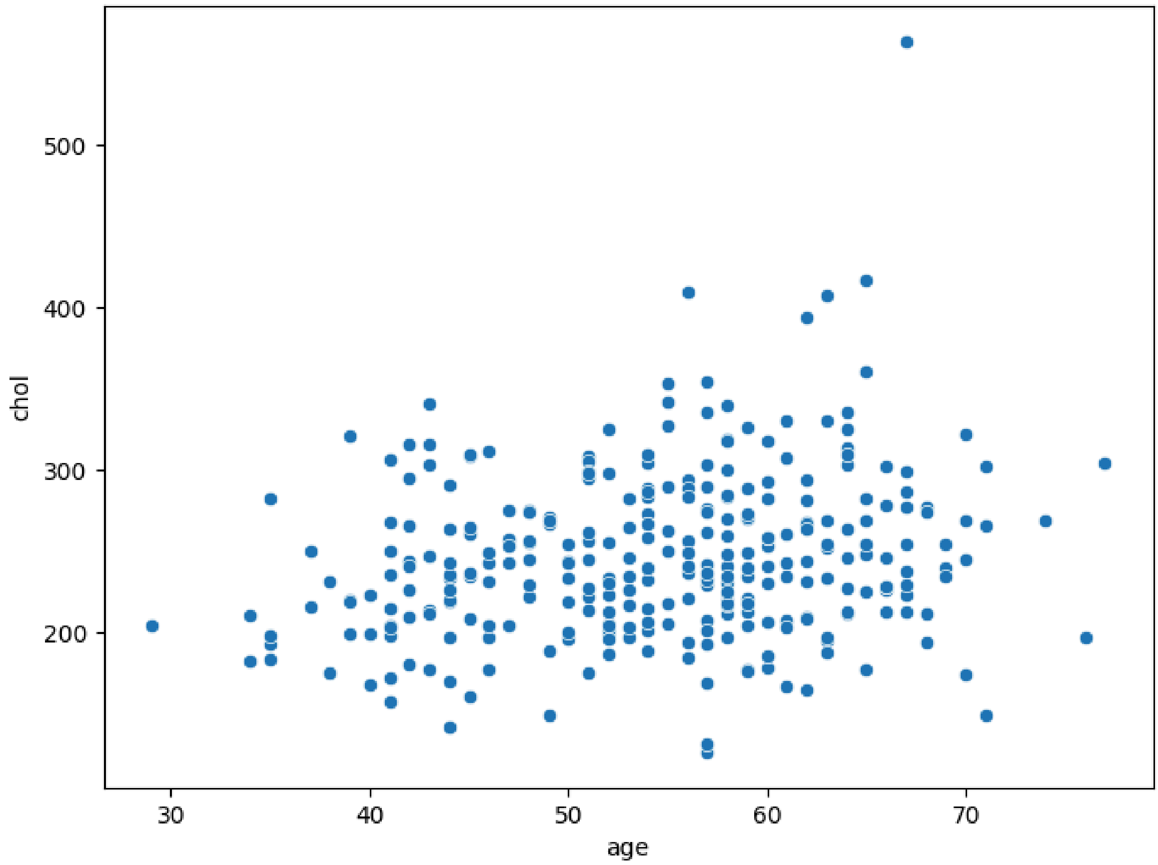
Interpretation

- The above line shows that linear regression model is not good fit to the data.

Analyze `age` and `chol` variable

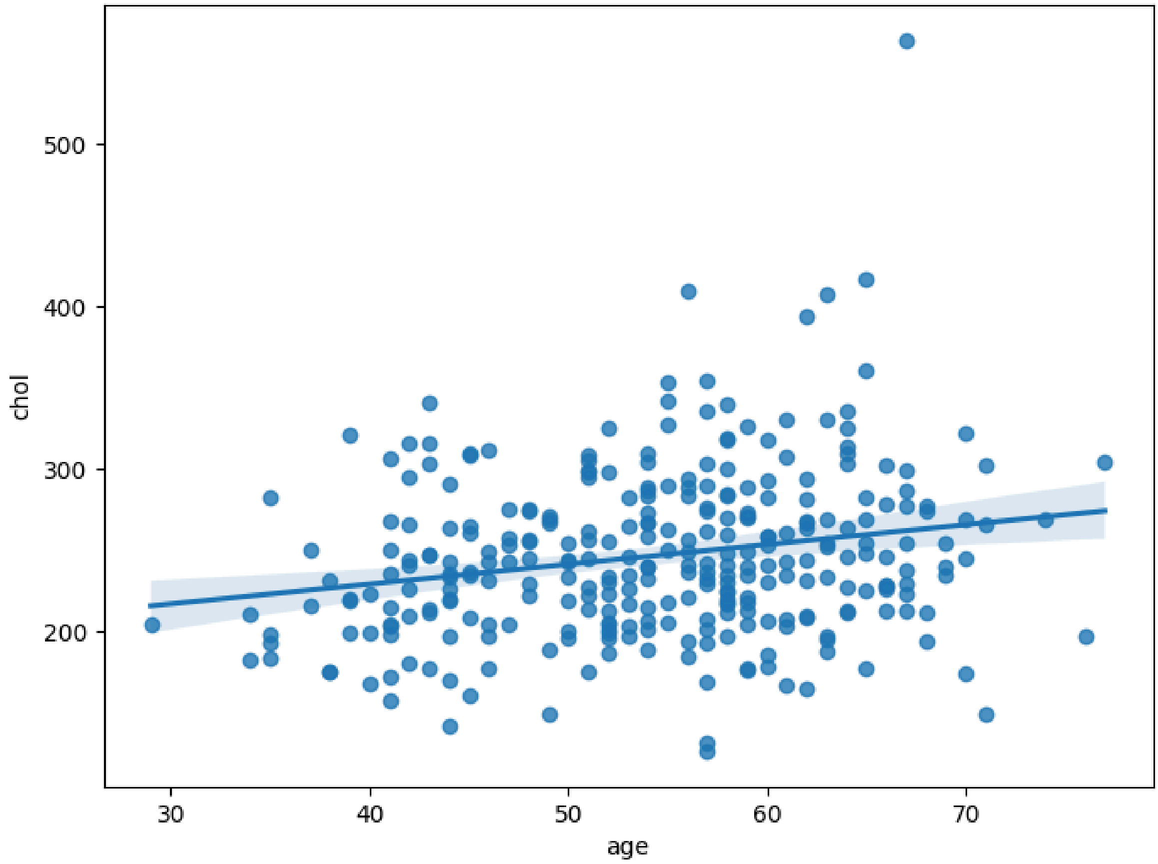
In [173…

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.scatterplot(x="age", y="chol", data=df)
plt.show()
```



In [175…

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.regplot(x="age", y="chol", data=df)
plt.show()
```



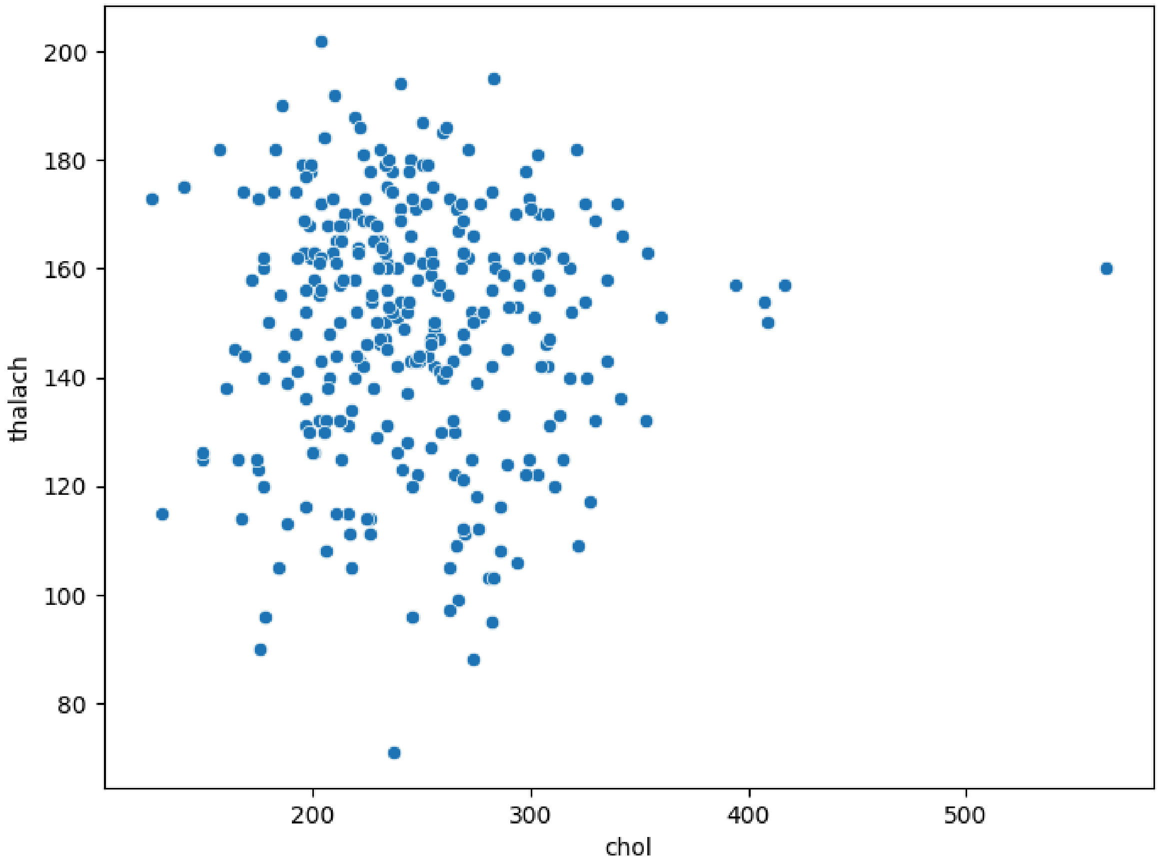
Interpretation

- The above plot confirms that there is a slighly positive correlation between `age` and `chol` variables.

Analyze `chol` and `thalach` variable

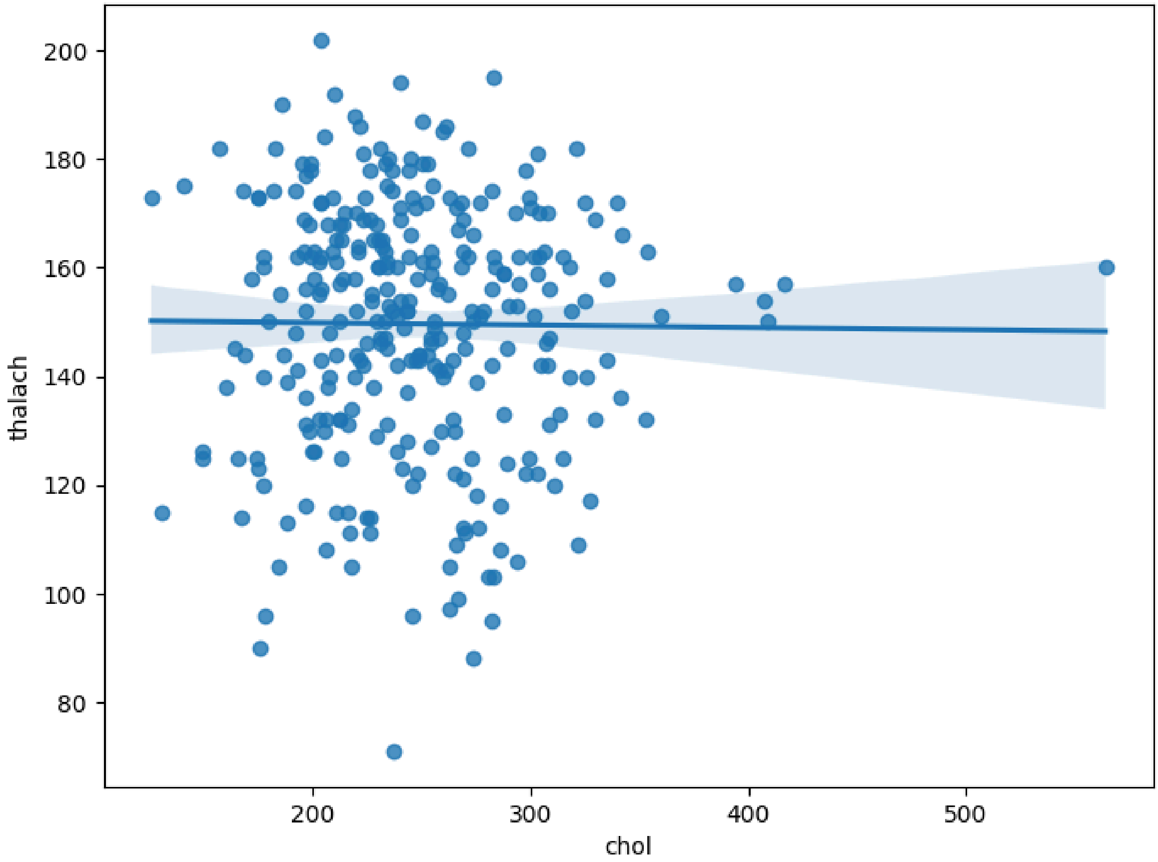
In [179…

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.scatterplot(x="chol", y = "thalach", data=df)
plt.show()
```



In [181…

```
f, ax = plt.subplots(figsize=(8, 6))
ax = sns.regplot(x="chol", y="thalach", data=df)
plt.show()
```



Dealing with missing values

- In Pandas missing data is represented by two values:
  - **None**: None is a Python singleton object that is often used for missing data in Python code.
  - **NaN** : NaN (an acronym for Not a Number), is a special floating-point value recognized by all systems that use the standard IEEE floating-point representation.
- There are different methods in place on how to detect missing values.

Pandas isnull() and notnull() functions

- Pandas offers two functions to test for missing data - `isnull()` and `notnull()`. These are simple functions that return a boolean value indicating whether the passed in argument value is in fact missing data.
- Below, I will list some useful commands to deal with missing values.

Useful commands to detect missing values

- **df.isnull()**

The above command checks whether each cell in a dataframe contains missing values or not. If the cell contains missing value, it returns True otherwise it returns False.

- **df.isnull().sum()**

The above command returns total number of missing values in each column in the dataframe.

- **df.isnull().sum().sum()**

It returns total number of missing values in the dataframe.

- **df.isnull().mean()**

It returns percentage of missing values in each column in the dataframe.

- **df.isnull().any()**

It checks which column has null values and which has not. The columns which has null values returns TRUE and FALSE otherwise.

- **df.isnull().any().any()**

It returns a boolean value indicating whether the dataframe has missing values or not. If dataframe contains missing values it returns TRUE and FALSE otherwise.

- **df.isnull().values.any()**

It checks whether a particular column has missing values or not. If the column contains missing values, then it returns TRUE otherwise FALSE.

- **df.isnull().values.sum()**

It returns the total number of missing values in the dataframe.

In [185…

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

Out[185…

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

Interpretation

We can see that there are no missing values in the dataset.

Check with ASSERT statement

- We must confirm that our dataset has no missing values.
- We can write an **assert statement** to verify this.
- We can use an assert statement to programmatically check that no missing, unexpected 0 or negative values are present.
- This gives us confidence that our code is running properly.
- **Assert statement** will return nothing if the value being tested is true and will throw an AssertionError if the value is false.
- **Asserts**
    - assert 1 == 1 (return Nothing if the value is True)
    - assert 1 == 2 (return AssertionError if the value is False)

In [192…

```
assert pd.notnull(df).all().all()
#assert that there are no missing values in the dataframe
```

In [194…

```
#assert all values are greater than or equal to 0

assert (df >= 0).all().all()
```

Outlier detection

I will make boxplots to visualise outliers in the continuous numerical variables : -

`age` , `trestbps` , `chol` , `thalach` and `oldpeak` variables.

`age` variable

In [205…

```
df['age'].describe()
```

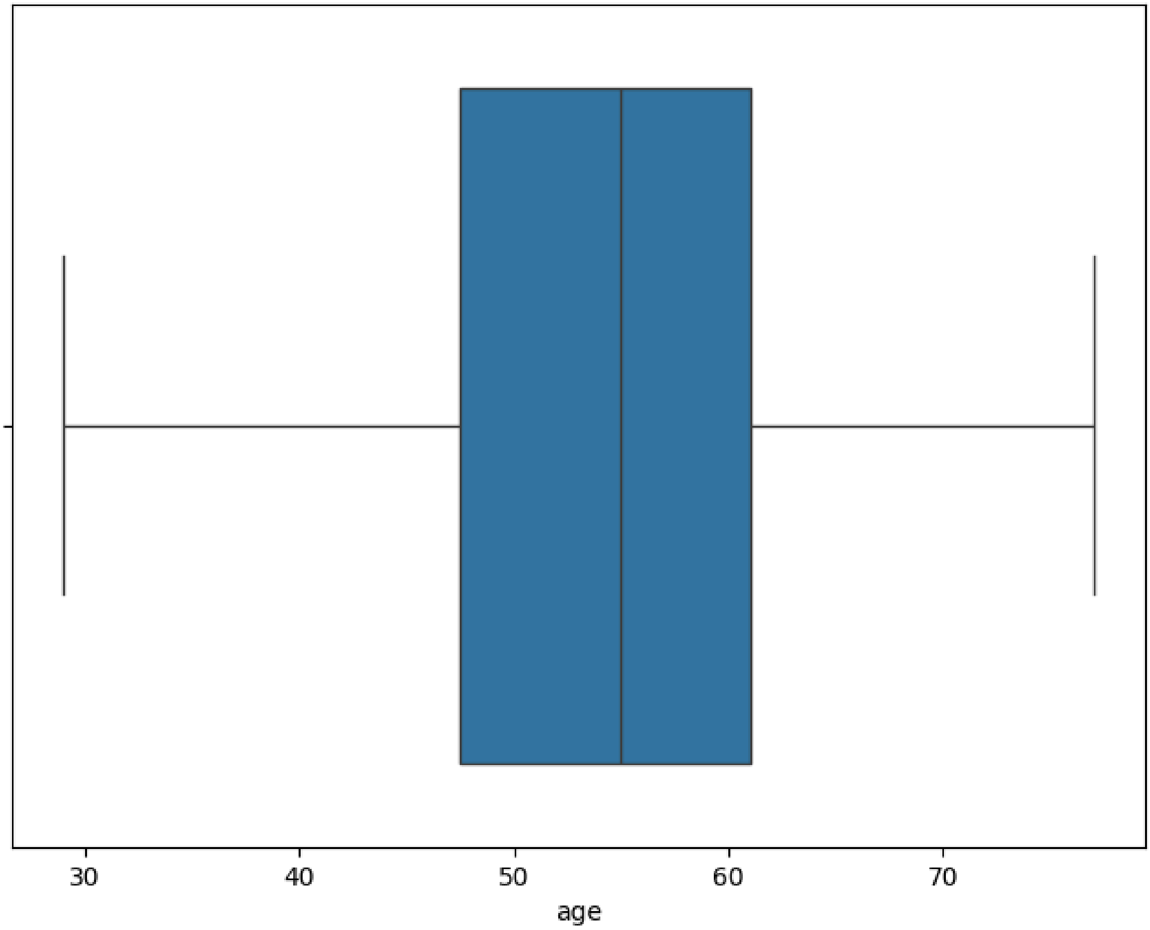
Out[205…

```
count    303.000000
mean      54.366337
std        9.082101
min       29.000000
25%       47.500000
50%       55.000000
75%       61.000000
max       77.000000
Name: age, dtype: float64
```

Box-plot of `age` variable

In [208…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x=df["age"])
plt.show()
```



`trestbps` variable

In [211…

```
df['trestbps'].describe()
```

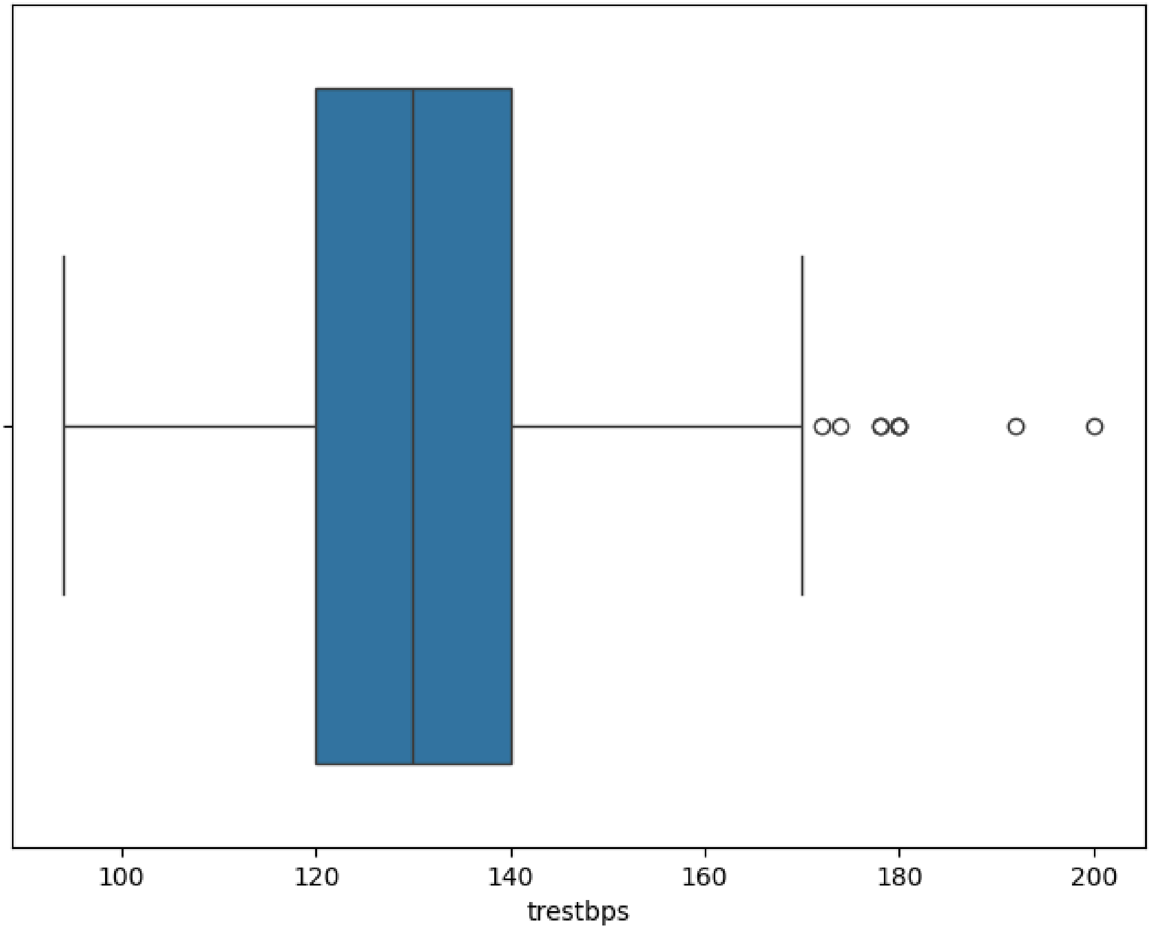
Out[211…

```
count    303.000000
mean     131.623762
std       17.538143
min       94.000000
25%      120.000000
50%      130.000000
75%      140.000000
max      200.000000
Name: trestbps, dtype: float64
```

Box-plot of `trestbps` variable

In [214…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x=df["trestbps"])
plt.show()
```



`chol` variable

In [217…

```
df['chol'].describe()
```

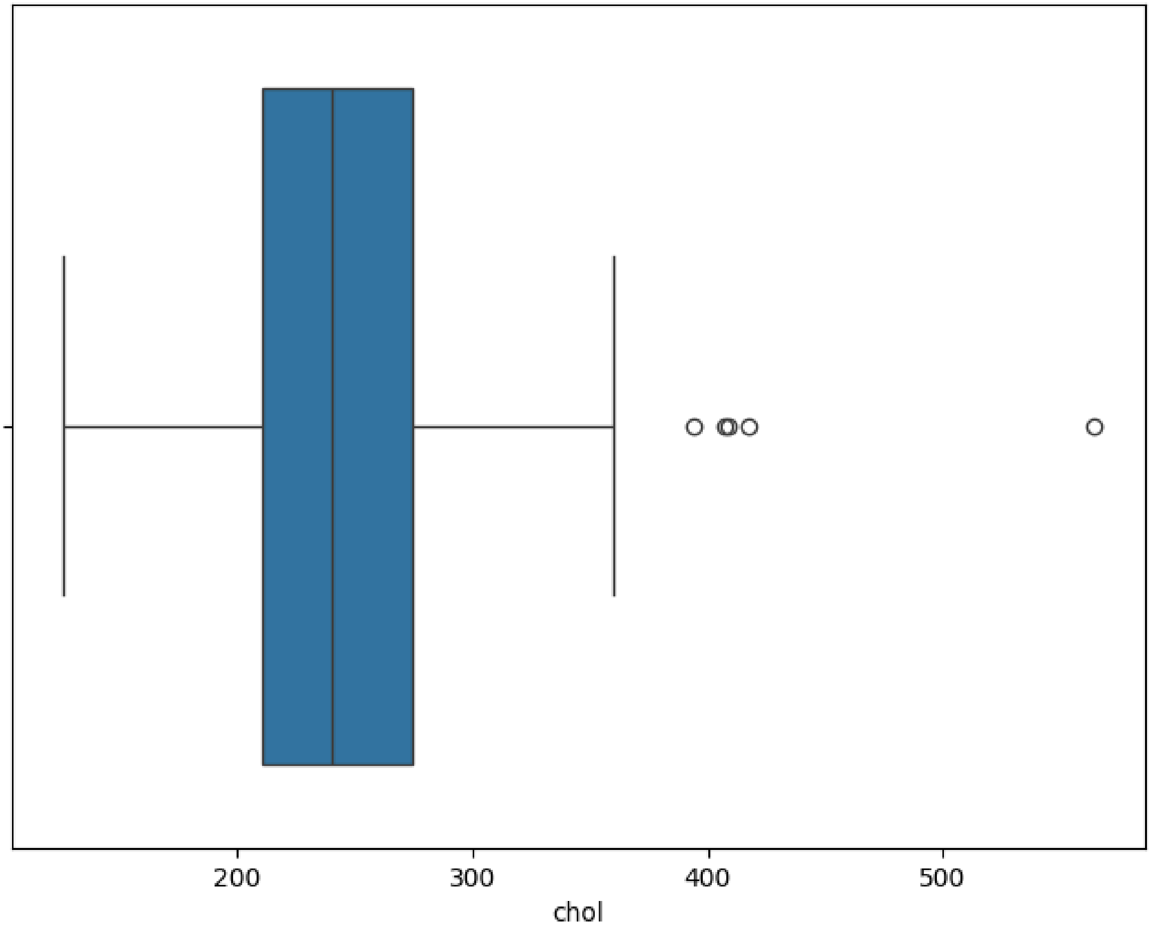
Out[217…

```
count    303.000000
mean     246.264026
std       51.830751
min      126.000000
25%      211.000000
50%      240.000000
75%      274.500000
max      564.000000
Name: chol, dtype: float64
```

Box-plot of `chol` variable

In [220…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x=df["chol"])
plt.show()
```



`thalach` variable

In [223…

```
df['thalach'].describe()
```

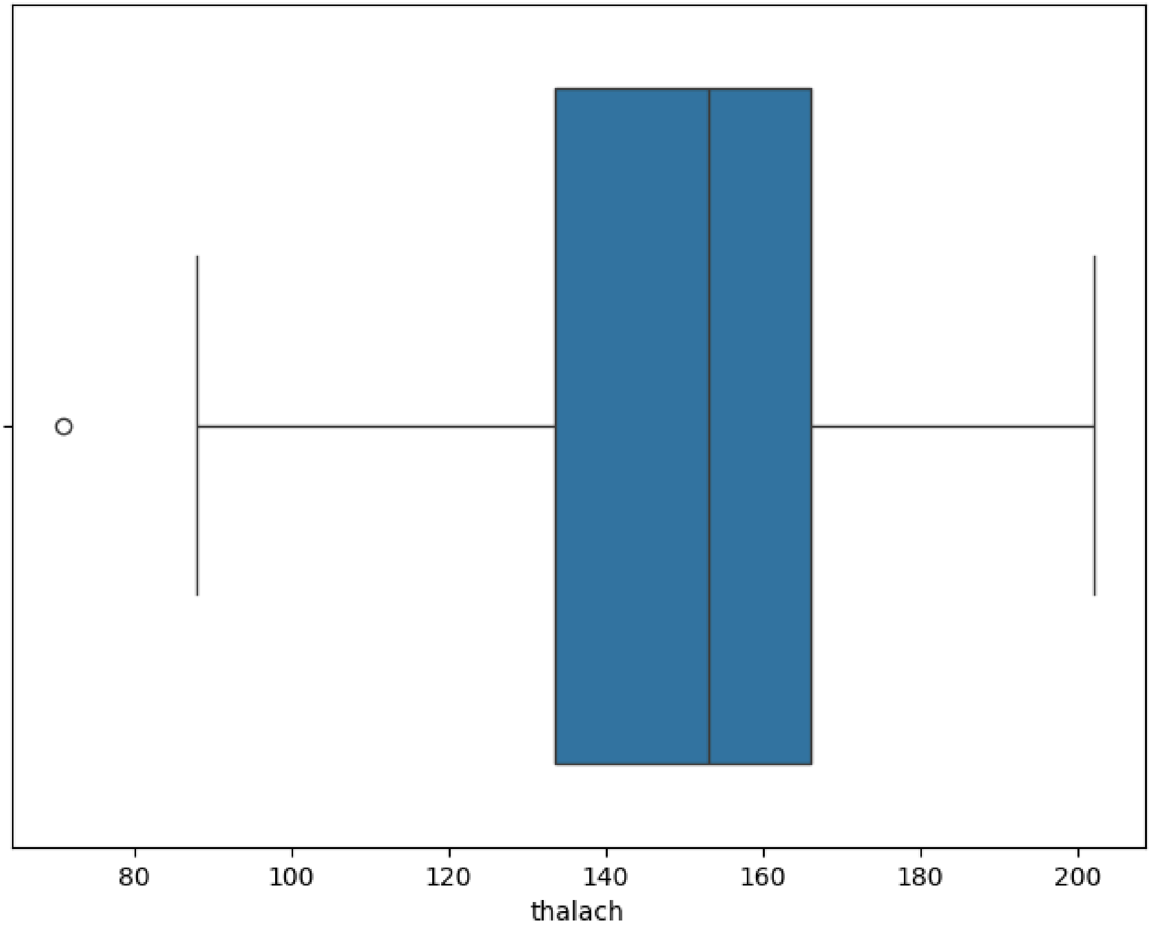
Out[223…

```
count    303.000000
mean     149.646865
std       22.905161
min       71.000000
25%      133.500000
50%      153.000000
75%      166.000000
max      202.000000
Name: thalach, dtype: float64
```

Box-plot of `thalach` variable

In [226…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x=df["thalach"])
plt.show()
```



`oldpeak` variable

In [229…

```
df['oldpeak'].describe()
```

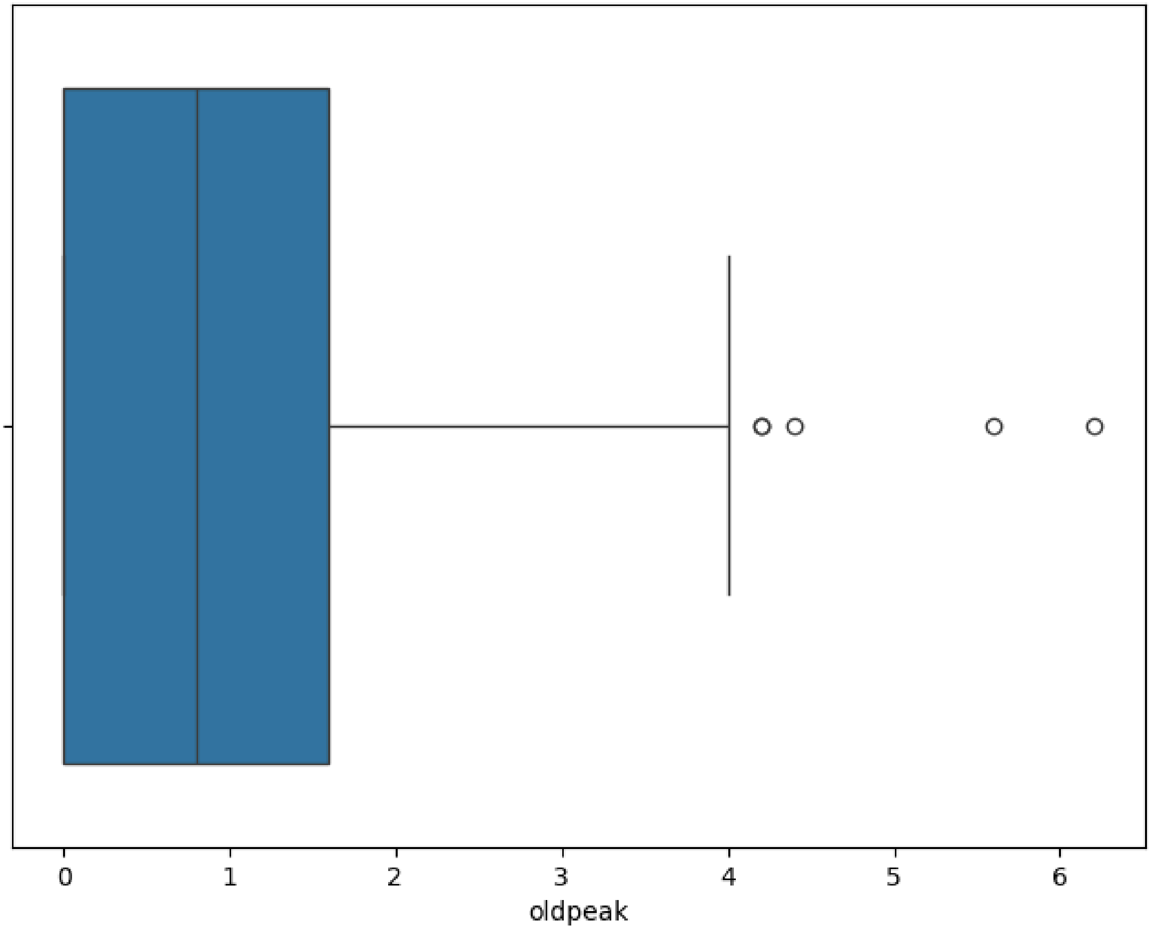
Out[229…

```
count    303.000000
mean       1.039604
std        1.161075
min        0.000000
25%        0.000000
50%        0.800000
75%        1.600000
max        6.200000
Name: oldpeak, dtype: float64
```

Box-plot of `oldpeak` variable

In [232…

```
f, ax = plt.subplots(figsize=(8, 6))
sns.boxplot(x=df["oldpeak"])
plt.show()
```



Findings

- The `age` variable does not contain any outlier.
- `trestbps` variable contains outliers to the right side.
- `chol` variable also contains outliers to the right side.
- `thalach` variable contains a single outlier to the left side.
- `oldpeak` variable contains outliers to the right side.
- Those variables containing outliers needs further investigation.

Conclusion

So, friends, our EDA journey has come to an end.

In this kernel, we have explored the heart disease dataset. In this kernel, we have implemented many of the strategies presented in the book Think Stats - Exploratory Data Analysis in Python by Allen B Downey . The feature variable of interest is target variable. We have analyzed it alone and check its interaction with other variables. We have also discussed how to detect missing data and outliers.

I hope you like this kernel on EDA journey.

Thanks

References

The following references are used to create this kernel

- Think Stats - Exploratory Data Analysis in Python by Allen B Downey
- Seaborn API reference
- My other kernel

In [ ]: