Python Lab Exercise #2

Objectives:

- Load .csv files into `pandas` DataFrames
- Describe and manipulate data in Series and DataFrames
- Visualize data using DataFrame methods and `matplotlib`

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

What is Pandas?

Pandas, as the Anaconda docs tell us, offers us "High-performance, easy-to-use data structures and data analysis tools." It's something like "Excel for Python", but it's quite a bit more powerful.

Let's read in the heart dataset.

Pandas has many methods for reading different types of files. Note that here we have a .csv file.

Read about this dataset here.

In [2]:
```
heart_df = pd.read_csv('/Users/linw/Desktop/DS311-Technologies-in-Data-Analytic-FA23/Wee
```

The output of the `.read_csv()` function is a pandas *DataFrame*, which has a familiar tabaular structure of rows and columns.

In [3]:
```
type(heart_df)
```

Out[3]:
```
pandas.core.frame.DataFrame
```

In [4]:
```
heart_df
```

Out[4]:

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

303 rows × 14 columns

DataFrames and Series

Two main types of pandas objects are the DataFrame and the Series, the latter being in effect a single column of the former:

In [5]:
```
age_series = heart_df['age']
type(age_series)
```

Out[5]:
```
pandas.core.series.Series
```

Notice how we can isolate a column of our DataFrame simply by using square brackets together with the name of the column.

Both Series and DataFrames have an *index* as well:

In [6]:
```
heart_df.index
```

Out[6]:
```
RangeIndex(start=0, stop=303, step=1)
```

In [7]:
```
age_series.index
```

Out[7]:
```
RangeIndex(start=0, stop=303, step=1)
```

Pandas is built on top of NumPy, and we can always access the NumPy array underlying a DataFrame using `.values`.

In [8]:
```
heart_df.values
```

Out[8]:
```
array([[63.,  1.,  3., ...,  0.,  1.,  1.],
       [37.,  1.,  2., ...,  0.,  2.,  1.],
       [41.,  0.,  1., ...,  0.,  2.,  1.],
       ...,
       [68.,  1.,  0., ...,  2.,  3.,  0.],
       [57.,  1.,  0., ...,  1.,  3.,  0.],
       [57.,  0.,  1., ...,  1.,  2.,  0.]])
```

Basic DataFrame Attributes and Methods

.head()

In [9]:
```
heart_df.head()
```

Out[9]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	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	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

.tail()

In [10]:
```
heart_df.tail()
```

Out[10]:

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

.info()

In [11]:
```
heart_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
```

`.describe()`

In [12]:
```
heart_df.describe()
```

Out[12]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000

`.dtypes`

In [13]:
```
heart_df.dtypes
```

Out[13]:
```
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
```

`.shape`

In [14]:
```
heart_df.shape
```

Out[14]:
```
(303, 14)
```

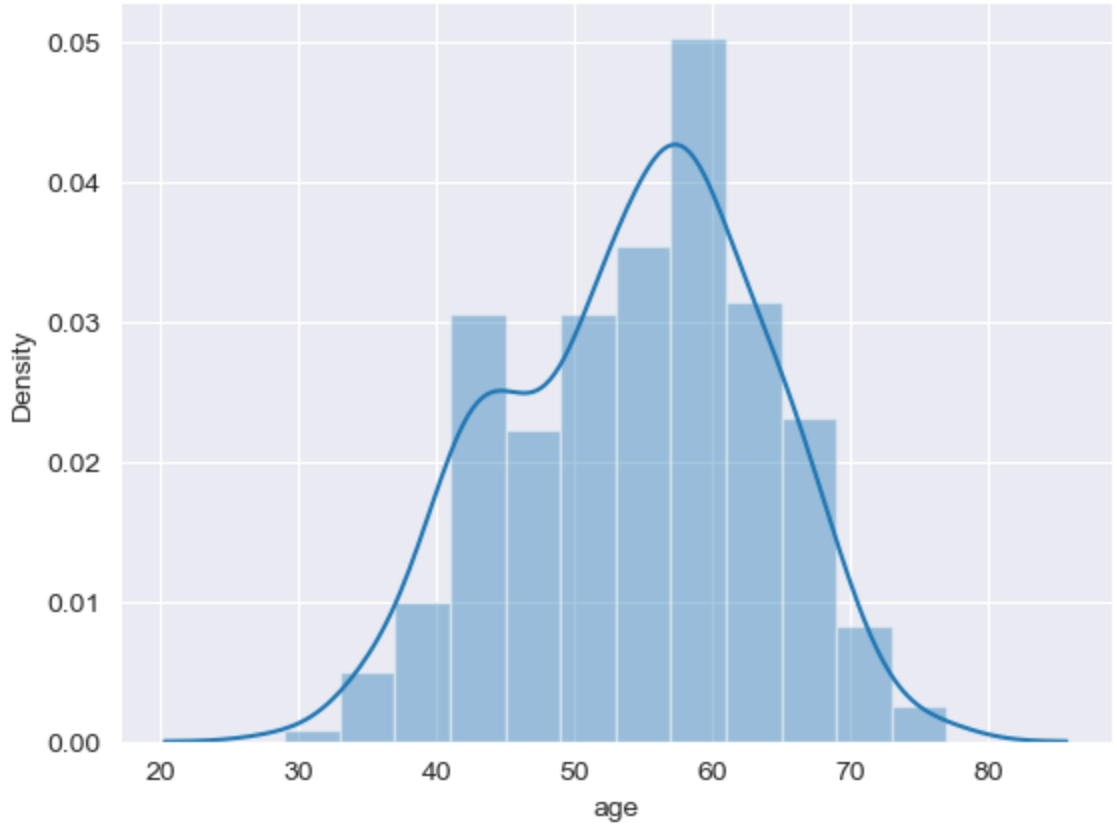
Exploratory Plots

Let's make ourselves a histogram of ages:

In [15]:
```
sns.set_style('darkgrid')
sns.distplot(a=heart_df['age']);
# For more recent versions of seaborn:
# sns.histplot(data=heart_df['age'], kde=True);
```

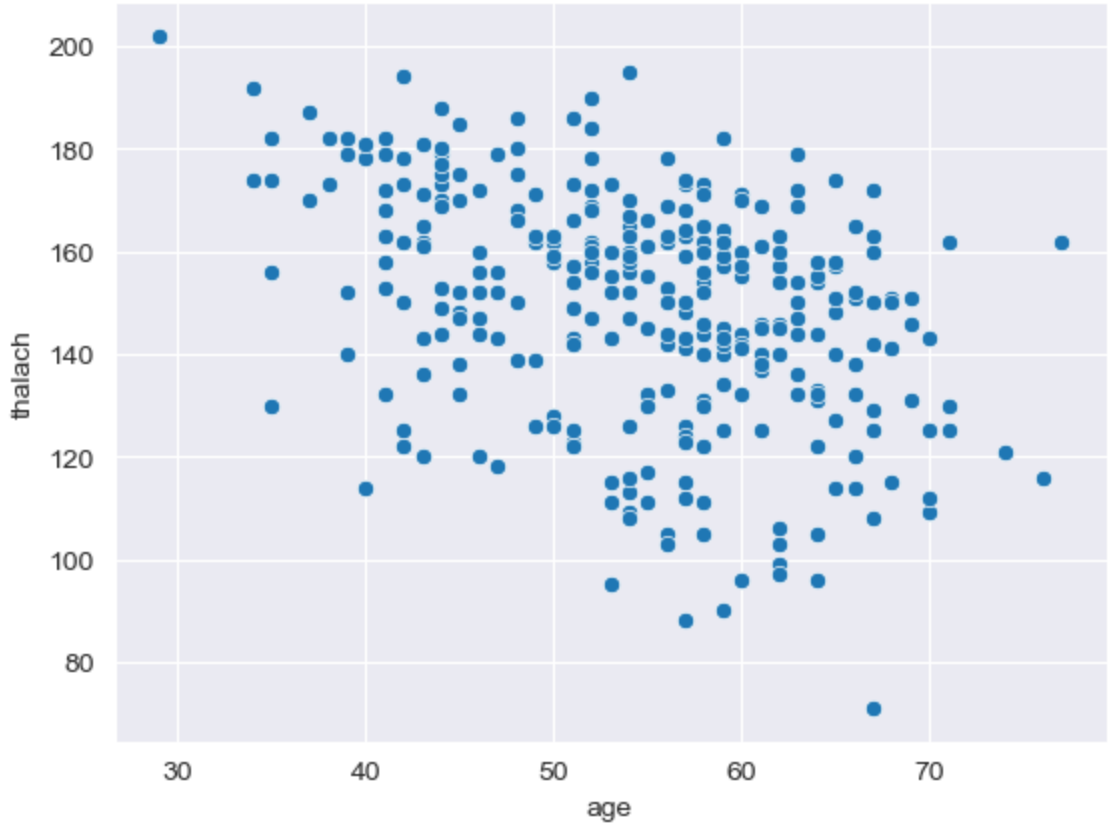
```
/var/folders/my/1fd3hsfs2gj1zs1vyd2xwpcc0000gn/T/ipykernel_86443/2016466243.py:2: UserWa
rning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.
```

```
Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(a=heart_df['age']);
```



And while we're at it let's do a scatter plot of maximum heart rate vs. age:

In [16]:
```
sns.scatterplot(x=heart_df['age'], y=heart_df['thalach']);
```



Adding to a DataFrame

Adding Rows

Here are two rows that our engineer accidentally left out of the .csv file, expressed as a Python dictionary:

In [17]:
```
extra_rows = {'age': [40, 30], 'sex': [1, 0], 'cp': [0, 0], 'trestbps': [120, 130],
              'chol': [240, 200],
             'fbs': [0, 0], 'restecg': [1, 0], 'thalach': [120, 122], 'exang': [0, 1],
              'oldpeak': [0.1, 1.0], 'slope': [1, 1], 'ca': [0, 1], 'thal': [2, 3],
              'target': [0, 0]}
extra_rows
```

Out[17]:
```
{'age': [40, 30],
 'sex': [1, 0],
 'cp': [0, 0],
 'trestbps': [120, 130],
 'chol': [240, 200],
 'fbs': [0, 0],
 'restecg': [1, 0],
 'thalach': [120, 122],
 'exang': [0, 1],
 'oldpeak': [0.1, 1.0],
 'slope': [1, 1],
 'ca': [0, 1],
 'thal': [2, 3],
 'target': [0, 0]}
```

How can we add this to the bottom of our dataset?

In [18]:
```
# Let's first turn this into a DataFrame.
# We can use the .from_dict() method.
```

```
missing = pd.DataFrame(extra_rows)
missing
```

Out[18]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	40	1	0	120	240	0	1	120	0	0.1	1	0	2	0
1	30	0	0	130	200	0	0	122	1	1.0	1	1	3	0

In [19]:
```
# Now we just need to concatenate the two DataFrames together.
# Note the `ignore_index` parameter! We'll set that to True.

heart_augmented = pd.concat([heart_df, missing],
                            ignore_index=True)
```

In [20]:
```
# Let's check the end to make sure we were successful!

heart_augmented.tail()
```

Out[20]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0
303	40	1	0	120	240	0	1	120	0	0.1	1	0	2	0
304	30	0	0	130	200	0	0	122	1	1.0	1	1	3	0

Adding Columns

Adding a column is very easy in `pandas`. Let's add a new column to our dataset called "test", and set all of its values to 0.

In [21]:
```
heart_augmented['test'] = 0
```

In [22]:
```
heart_augmented.head()
```

Out[22]:

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

I can also add columns whose values are functions of existing columns.

Suppose I want to add the cholesterol column ("chol") to the resting systolic blood pressure column ("trestbps"):

In [23]:
```
heart_augmented['chol+trestbps'] = heart_augmented['chol'] + heart_augmented['trestbps']
```

In [24]:
```
heart_augmented.head()
```

Out[24]:

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

Filtering

We can use filtering techniques to see only certain rows of our data. If we wanted to see only the rows for patients 70 years of age or older, we can simply type:

In [25]:
```
heart_augmented['age'] >= 70
```

Out[25]:
```
0      False
1      False
2      False
3      False
4      False
       ...
300    False
301    False
302    False
303    False
304    False
Name: age, Length: 305, dtype: bool
```

In [26]:
```
heart_augmented[heart_augmented['age'] >= 70]
```

Out[26]:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	test	chol
25	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1	0	
60	71	0	2	110	265	1	0	130	0	0.0	2	1	2	1	0	
129	74	0	1	120	269	0	0	121	1	0.2	2	1	2	1	0	
144	76	0	2	140	197	0	2	116	0	1.1	1	0	2	1	0	
145	70	1	1	156	245	0	0	143	0	0.0	2	0	2	1	0	
151	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1	0	
225	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0	0	
234	70	1	0	130	322	0	0	109	0	2.4	1	3	2	0	0	
238	77	1	0	125	304	0	0	162	1	0.0	2	3	2	0	0	
240	70	1	2	160	269	0	1	112	1	2.9	1	1	3	0	0	

Use '&' for "and" and '|' for "or".

Exercise

Display the patients who are 70 or over as well as the patients whose trestbps score is greater than 170.

In [27]:
```
# Enter your code here
heart_augmented[(heart_augmented['age'] >= 70) | (heart_augmented['trestbps'] > 170)]
```

Out[27]:

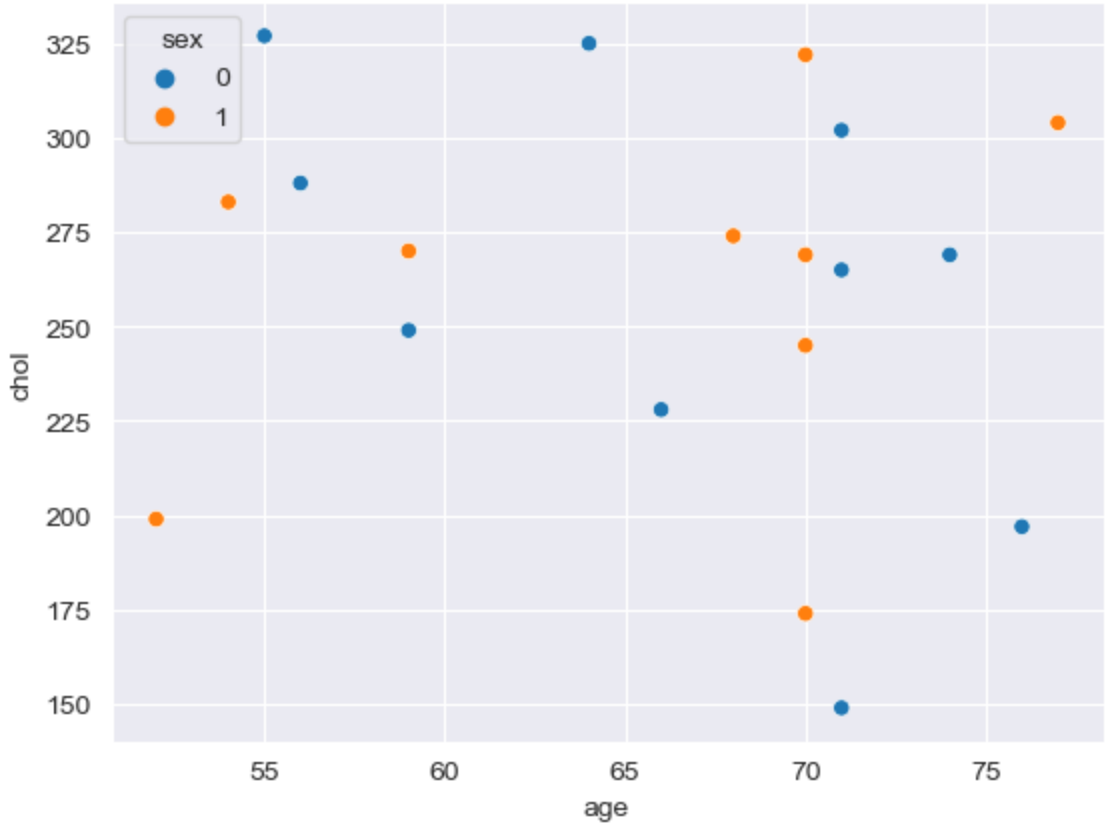
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	test	chol
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1	0	
25	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1	0	
60	71	0	2	110	265	1	0	130	0	0.0	2	1	2	1	0	
101	59	1	3	178	270	0	0	145	0	4.2	0	0	3	1	0	
110	64	0	0	180	325	0	1	154	1	0.0	2	0	2	1	0	
129	74	0	1	120	269	0	0	121	1	0.2	2	1	2	1	0	
144	76	0	2	140	197	0	2	116	0	1.1	1	0	2	1	0	
145	70	1	1	156	245	0	0	143	0	0.0	2	0	2	1	0	
151	71	0	0	112	149	0	1	125	0	1.6	1	0	2	1	0	
203	68	1	2	180	274	1	0	150	1	1.6	1	0	3	0	0	
223	56	0	0	200	288	1	0	133	1	4.0	0	2	3	0	0	
225	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0	0	
234	70	1	0	130	322	0	0	109	0	2.4	1	3	2	0	0	
238	77	1	0	125	304	0	0	162	1	0.0	2	3	2	0	0	
240	70	1	2	160	269	0	1	112	1	2.9	1	1	3	0	0	
241	59	0	0	174	249	0	1	143	1	0.0	1	0	2	0	0	
248	54	1	1	192	283	0	0	195	0	0.0	2	1	3	0	0	
260	66	0	0	178	228	1	1	165	1	1.0	1	2	3	0	0	
266	55	0	0	180	327	0	2	117	1	3.4	1	0	2	0	0	

Exploratory Plot

Using the subframe we just made, let's make a scatter plot of their cholesterol levels vs. age and color by sex:

In [28]:

```
at_risk = heart_augmented[(heart_augmented['age'] >= 70) | (heart_augmented['trestbps']

sns.scatterplot(data=at_risk, x='age', y='chol', hue='sex');
```



`.loc` and `.iloc`

We can use `.loc` to get, say, the first ten values of the age and resting blood pressure ("trestbps") columns:

In [29]:
```
heart_augmented.loc
```

Out[29]:
```
<pandas.core.indexing._LocIndexer at 0x12f4ca3a0>
```

In [30]:
```
heart_augmented.loc[:9, ['age', 'trestbps']]
```

Out[30]:

	age	trestbps
0	63	145
1	37	130
2	41	130
3	56	120
4	57	120
5	57	140
6	56	140
7	44	120
8	52	172
9	57	150

`.iloc` is used for selecting locations in the DataFrame **by number**:

In [31]:
```
heart_augmented.iloc
```

Out[31]:
```
<pandas.core.indexing._iLocIndexer at 0x12f5667b0>
```

In [32]:
```
heart_augmented.iloc[3, 0]
```

Out[32]:
```
56
```

In [33]:
```
heart_augmented.head()
```

Out[33]:

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

Exercise

How would we get the same slice as just above by using .iloc() instead of .loc()?

In [34]:
```
# Enter your code here
heart_augmented.iloc[0:10, [0, 3]]
```

Out[34]:

	age	trestbps
0	63	145
1	37	130
2	41	130
3	56	120
4	57	120
5	57	140
6	56	140
7	44	120
8	52	172
9	57	150

Statistics

`.mean()`

In [35]:
```
heart_augmented.mean()
```

Out[35]:
```
age          54.239344
sex           0.681967
cp            0.960656
trestbps    131.580328
chol        246.091803
fbs           0.147541
restecg       0.527869
```

```
thalach          149.459016
exang              0.327869
oldpeak            1.036393
slope              1.396721
ca                 0.727869
thal               2.314754
target             0.540984
test               0.000000
chol+trestbps    377.672131
dtype: float64
```

Be careful! Some of these will are not straightforwardly interpretable. What does an average "sex" of 0.682 mean?

`.min()`

In [36]:
```
heart_augmented.min()
```

Out[36]:
```
age               29.0
sex                0.0
cp                 0.0
trestbps          94.0
chol             126.0
fbs                0.0
restecg            0.0
thalach           71.0
exang              0.0
oldpeak            0.0
slope              0.0
ca                 0.0
thal               0.0
target             0.0
test               0.0
chol+trestbps    249.0
dtype: float64
```

`.max()`

In [37]:
```
heart_augmented.max()
```

Out[37]:
```
age               77.0
sex                1.0
cp                 3.0
trestbps         200.0
chol             564.0
fbs                1.0
restecg            2.0
thalach          202.0
exang              1.0
oldpeak            6.2
slope              2.0
ca                 4.0
thal               3.0
target             1.0
test               0.0
chol+trestbps    679.0
dtype: float64
```

Series Methods

`.value_counts()`

How many different values does slope have? What about sex? And target?

In [38]:
```
heart_augmented['slope'].value_counts()
```

Out[38]:
```
2    142
1    142
0     21
Name: slope, dtype: int64
```

In [39]:
```
heart_augmented['sex'].value_counts()
```

Out[39]:
```
1    208
0     97
Name: sex, dtype: int64
```

`.sort_values()`

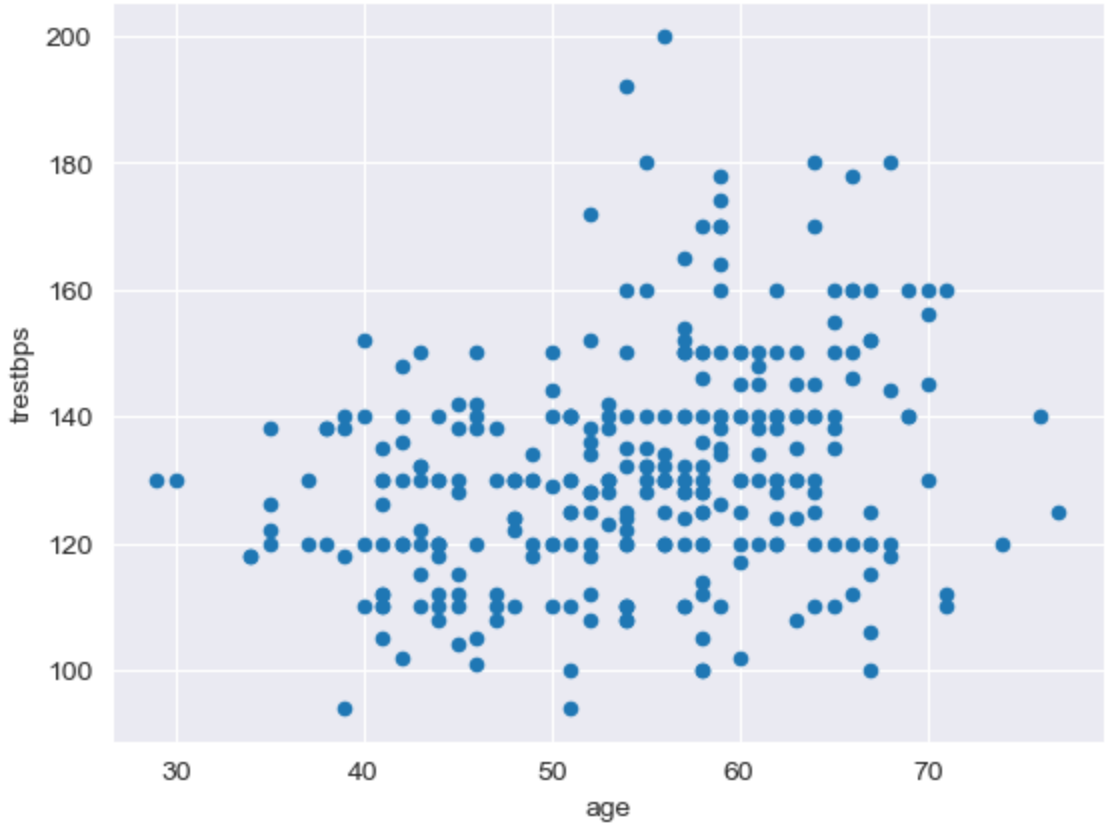
In [40]:
```
heart_augmented['age'].sort_values()
```

Out[40]:
```
72     29
304    30
58     34
125    34
65     35
       ..
25     71
60     71
129    74
144    76
238    77
Name: age, Length: 305, dtype: int64
```

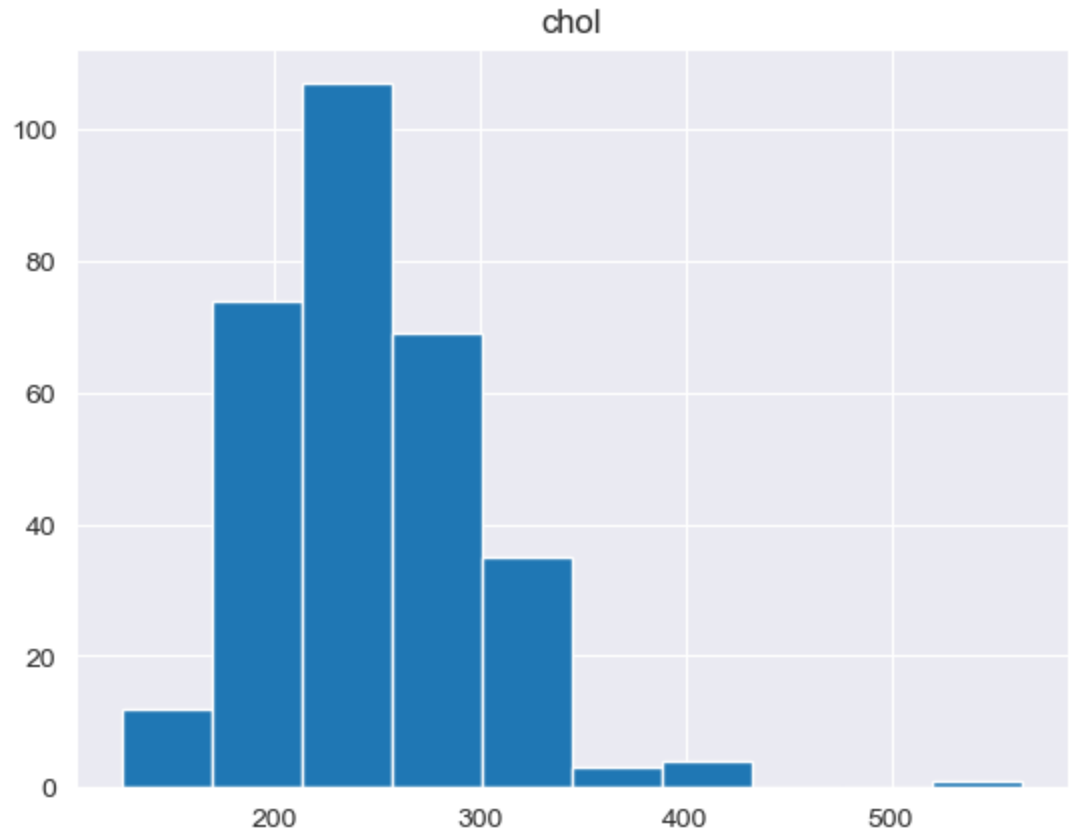
`pandas` -Native Plotting

The `.plot()` and `.hist()` methods available for DataFrames use a wrapper around `matplotlib`:

In [41]:
```
heart_augmented.plot(x='age', y='trestbps', kind='scatter');
```



In [42]:
```
heart_augmented.hist(column='chol');
```

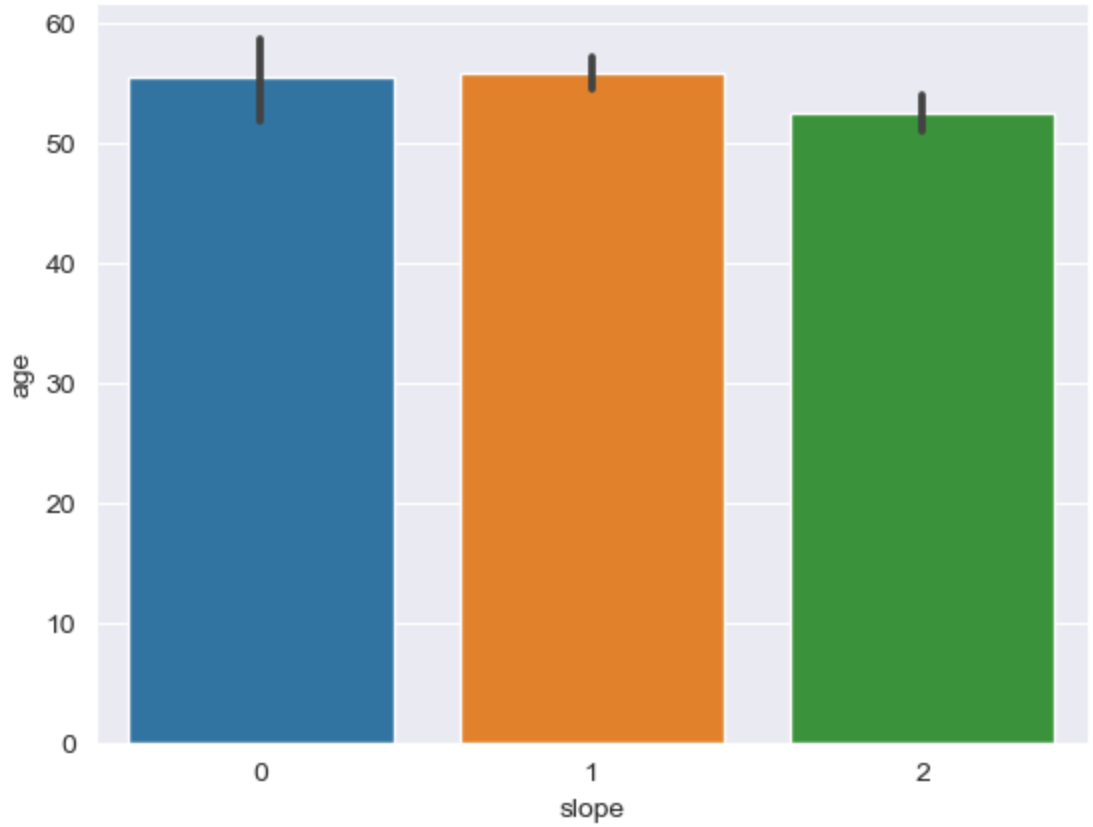


Exercises

1. Make a bar plot of "age" vs. "slope" for the `heart_augmented` DataFrame.

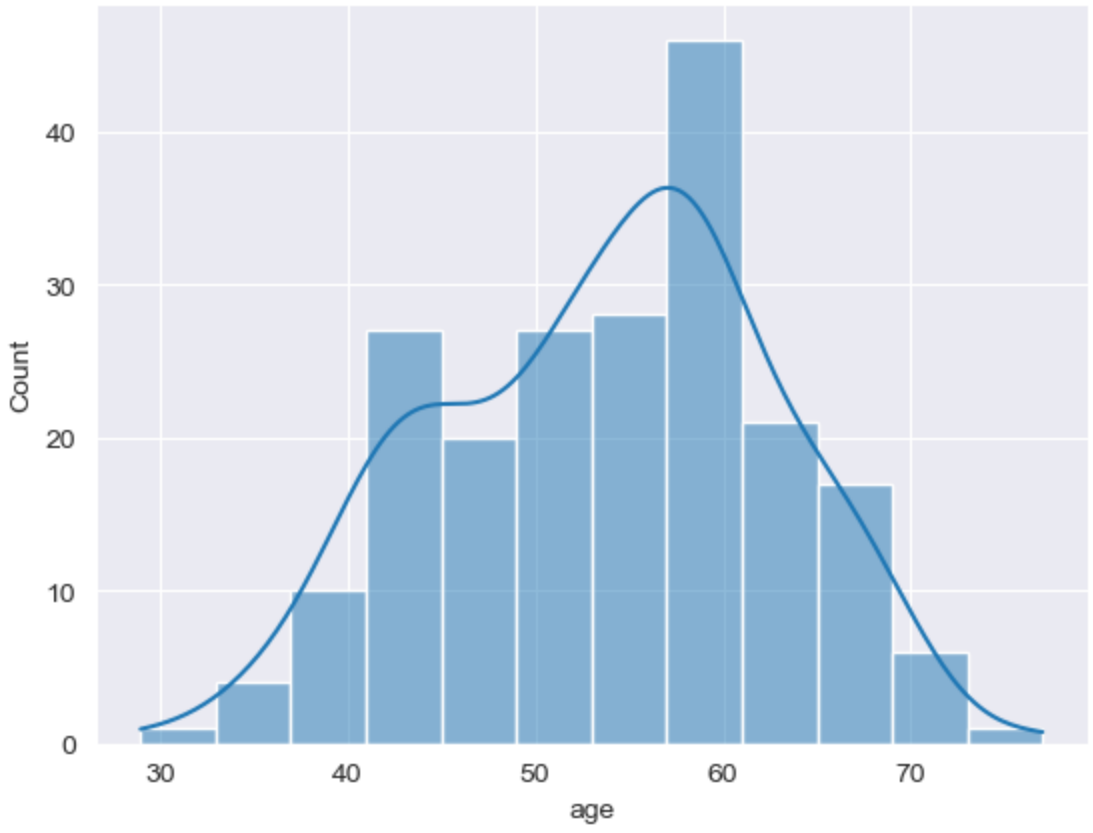
In [43]:
```
# Enter your code here
```

```
sns.barplot(data=heart_augmented, x='slope', y='age');
```



1. Make a histogram of ages for **just the men** in `heart_augmented` (heart_augmented['sex']=1).

In [44]:
```
# Enter your code here
men = heart_augmented[heart_augmented['sex'] == 1]
#sns.distplot(a=men['age']);
sns.histplot(data=men['age'], kde=True);
```



1. Make separate scatter plots of cholesterol vs. resting systolic blood pressure for the target=0 and the target=1 groups. Put both plots on the same figure and give each an appropriate title.

In [45]:
```
# Enter your code here
target0 = heart_augmented[heart_augmented['target'] == 0]
target1 = heart_augmented[heart_augmented['target'] == 1]
fig, ax = plt.subplots(1, 2, figsize=(11, 5), sharex=True, sharey=True)
sns.scatterplot(data=target0, x='trestbps', y='chol', ax=ax[0])
sns.scatterplot(data=target1, x='trestbps', y='chol', ax=ax[1])
ax[0].set_title('Cholesterol Vs. Resting Blood Pressure, No Heart Disease')
ax[1].set_title('Cholesterol Vs. Resting Blood Pressure, Heart Disease');
```