Data Wrangling

Data wrangling generally refers to the process of getting a data set ready for analysis. Why would we need to do that?

Real-world data can be messy. Data sets are recorded and assembled by humans, and humans make mistakes. A single data set might created and updated by multiple people who may decide to do things in slightly different ways. On a spreadsheet, one person might might decide to leave cells with missing data blank, another might enter "NaN", while a third may enter "missing". If the data has many many rows, one person might decide to repeat the column headers partway down so they don't have to scroll up to see them. Any of these things mean that the data set cannot be analyzed "as is" and wrangling will be required.

Even in a tightly controlled laboratory setting in which data are collected via computer and automatically written out to data files, some data wrangling might be required. There might be a separate data file for each subject or experimental session, meaning that these separate files will have to be combined into a single data set before analysis.

Our main wrangling tool is pandas, so we can go ahead and import it.

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

Loading

For our wrangling practice today, we'll look at a data set containing various measurements on breast cancer patients. The file is called `breast_cancer_data.csv` , and you should place it in the "data" folder you should already have in the same directory as this notebook.

Let's import it as a pandas dataframe.

In [2]:
```
bcd = pd.read_csv('./breast_cancer_data.csv')
bcd
```

Out[2]:

ell_size_uniformity	cell_shape_uniformity	marginal_adhesion	single_ep_cell_size	bare_nuclei	blan
1.0	1	1	2	1	
4.0	4	5	7	10	
1.0	1	1	2	2	
8.0	8	1	3	4	
1.0	1	3	2	1	
...	...	...	...	...	
1.0	1	1	3	2	
1.0	1	1	2	1	
10.0	10	3	7	3	
8.0	6	4	3	4	
8.0	8	5	4	5	

Before we do any actual wrangling, let's get familiar with the data frame in its current form.

Exploring the Data Frame

We can explore the data frame by looking at it's attributes, such as its shape, column names, and data types:

In [3]:
```
bcd.columns
```

Out[3]:
```
Index(['patient_id', 'clump_thickness', 'cell_size_uniformity',
       'cell_shape_uniformity', 'marginal_adhesion', 'single_ep_cell_
size',
       'bare_nuclei', 'bland_chromatin', 'normal_nucleoli', 'mitoses'
, 'class',
       'doctor_name'],
      dtype='object')
```

Use the cells below to get the shape and data types (`dtypes`) of our data frame.

In [4]:
```
bcd.shape
```

Out[4]:
```
(699, 12)
```

In [5]:
```
bcd.dtypes
```

Out[5]:
```
patient_id                  int64
clump_thickness           float64
cell_size_uniformity      float64
cell_shape_uniformity       int64
marginal_adhesion           int64
single_ep_cell_size         int64
bare_nuclei                object
bland_chromatin           float64
normal_nucleoli           float64
mitoses                     int64
class                      object
doctor_name                object
dtype: object
```

In the cell below, use the `describe()` method to get a summary of the numerical columns.

In [6]:
```
bcd.describe()
```

Out[6]:

	patient_id	clump_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adh
count	6.990000e+02	698.000000	698.000000	699.000000	699.00
mean	1.071704e+06	4.416905	3.137536	3.207439	2.79
std	6.170957e+05	2.817673	3.052575	2.971913	2.84
min	6.163400e+04	1.000000	1.000000	1.000000	1.00
25%	8.706885e+05	2.000000	1.000000	1.000000	1.00
50%	1.171710e+06	4.000000	1.000000	1.000000	1.00
75%	1.238298e+06	6.000000	5.000000	5.000000	3.50
max	1.345435e+07	10.000000	10.000000	10.000000	10.00

Modifying a text column

We'll often want to "tune up" columns that contain text. We might encounter, for example, a column containing full names that we need to break up into separate columns for the first and last names.

Let's look at the column for the doctors' names. Use the cell below to take a peek.

In [15]:
```
doc_names = bcd.groupby('doctor_name')
doc_names
```

Out[15]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x11faa6dd0>
```

In [16]:
```
docs_names = bcd['doctor_name']
docs_names
```

Out[16]:
```
0        Dr. Doe
1      Dr. Smith
2        Dr. Lee
3      Dr. Smith
4       Dr. Wong
         ...
694      Dr. Lee
695    Dr. Smith
696      Dr. Lee
697      Dr. Lee
698     Dr. Wong
Name: doctor_name, Length: 699, dtype: object
```

The doctors' name data are redundant; each one has a "Dr. " in front of the actual name, but we already know these are doctors by the column name. Further, the entries have white space in them, which can cause us problems down the road. So let's modify this column so it only contains the surnames of the doctors.

One great thing about pandas is that it has versions of many of Python's string methods that operate *element-wise on an entire column of strings*. Here, we want to separate the "Dr. " from the actual name, which is exactly what Python's `str.split()` function does. So chances are, pandas has a version of this function that operates element-wise on data frames.

String Splitting Review:

Let's briefly remind ourselves of splitting up Python strings and extracting bits of them.

In [9]:
```
# Here's a string of the form: surname, first initial.
myStr = 'SirString, A.'
print(myStr)
```

```
SirString, A.
```

Let's say we wanted to get the surname. We could split this string into a Python list at the white space like this:

In [10]:
```
spltStr = myStr.split()    # split() defaults to splitting at white sp
print(spltStr)
```

```
['SirString,', 'A.']
```

We now have a list in which the items contain the text on either side of the split. This is close to what we want: the first entry in the list has the surname, but it also has an unwanted comma.

Let's split the string at the comma instead:

In [11]:
```
spltStr = myStr.split(',')   # tell Python to split at commas
print(spltStr)
```

```
['SirString', ' A.']
```

Now we have isolated the last name, and we can fetch it by indexing:

In [12]:
```
surname = spltStr[0]
print(surname)
```

```
SirString
```

In the cell below, see if you can extract the surname from `myStr` in one line of code:

In [20]:
```
surname = myStr.split(',')[0]
print(surname)
```

```
SirString
```

Alright, time to replace the `bcd['doctor_name']` column values with just the doctors' last names.

We could do this in one step, but let's break it out for clarity. First, let's copy the name column out into a new series.

In [13]:
```
dr_names = bcd['doctor_name']
dr_names
```

Out[13]:
```
0        Dr. Doe
1      Dr. Smith
2        Dr. Lee
3      Dr. Smith
4       Dr. Wong
         ...
694      Dr. Lee
695    Dr. Smith
696      Dr. Lee
697      Dr. Lee
698     Dr. Wong
Name: doctor_name, Length: 699, dtype: object
```

Note: pandas objects behave like ordinary Python objects. So, strictly speaking, we have not created a new object (pandas Series), rather, *we have created a new label that refers to the "doctor_name" column of `bcd` .*

In the cell below, use the `id()` function to compare the object IDs of `dr_names` and the corresponding column of `bcd` .

In [24]:
```
dr_names = bcd['doctor_name']
print("id dr_names:", id(dr_names))
print("id corresponding column of bcd:", id(bcd['doctor_name']))
```

```
dr_names: 4867126288
corresponding column of bcd: 4867126288
```

Now let's split all the names in the `doctor_name` column at the whitespace by using pandas `DataFrame.str.split()` function.

In [25]:
```
split_dr_names = dr_names.str.split()
split_dr_names
```

Out[25]:
```
0      [Dr., Smith]
1    [Dr., Johnson]
2      [Dr., Patel]
Name: doctor_name, dtype: object
```

`DataFrame.str.split()` , however, *does* create a new object.

Use the cell below to confirm that the `split()` spawed a new object.

In [30]:
```
dr_names = bcd['doctor_name']
split_dr_names = dr_names.str.split()
print("id dr_names:", id(dr_names))
print("id split_dr_names:", id(split_dr_names))
```

```
id dr_names: 4867126288
id split_dr_names: 4867129040
```

Now we have a column of lists, each with two elements. The first element of each list is the "Dr. " bit, and the second consists of the surnames we want.

We can get these by using pandas string indexing, `Series.str[index]` .

In [26]:
```
surnames = split_dr_names.str[1]
surnames
```

Out[26]:
```
0      Smith
1    Johnson
2      Patel
Name: doctor_name, dtype: object
```

Note that, like the splitting, the string indexing worked on the entire `Series` automatically.

Now we can change the column in our main data frame, `bcd` .

In [31]:
```
bcd['doctor_name'] = surnames
```

In [32]:
```
bcd['doctor_name']
```

Out[32]:
```
0      Smith
1    Johnson
2      Patel
Name: doctor_name, dtype: object
```

Success!

Converting a column type (and other aggravations)

Let's look at those data types again.

In [33]:
```
bcd.dtypes
```

Out[33]:
```
doctor_name    object
dtype: object
```

Notice that "class" and "doctor_name" are of dtype "object", which refers to a general purpose column type, and is how pandas imports text columns by default. Most of the others are numeric (integers or floats), except for "bare_nuclei".

In the cell below, take a quick glance at 'bcd' again, and see if the "bare_nuclei" column should be a different data type than, say, the "marginal_adhesion" column.

In [38]:
```
bcd
```

Out[38]:

_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adhesion	single_ep_cell_size	bar
5.0	1.0	1	1	2	
5.0	4.0	4	5	7	
3.0	1.0	1	1	2	
6.0	8.0	8	1	3	
4.0	1.0	1	3	2	
...	...	...	...	...	
3.0	1.0	1	1	3	
2.0	1.0	1	1	2	
5.0	10.0	10	3	7	
4.0	8.0	6	4	3	
4.0	8.0	8	5	4	

s

It looks like "bare_nuclei" was intended to be a numeric column, so let's try and convert it using the `DataFrame.astype()` converter method.

In [39]:
```
bcd['bare_nuclei'] = bcd['bare_nuclei'].astype('int64')
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call
last)

Cell In[39], line 1
----> 1 bcd['bare_nuclei'] = bcd['bare_nuclei'].astype('int64')

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/generic.py:
6324, in NDFrame.astype(self, dtype, copy, errors)
   6317     results = [
   6318         self.iloc[:, i].astype(dtype, copy=copy)
   6319         for i in range(len(self.columns))
   6320     ]
   6322 else:
   6323     # else, only a single dtype is given
-> 6324     new_data = self._mgr.astype(dtype=dtype, copy=copy, error
s=errors)
   6325     return self._constructor(new_data).__finalize__(self, met
hod="astype")
   6327 # GH 33113: handle empty frame or series

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/m
anagers.py:451, in BaseBlockManager.astype(self, dtype, copy, errors)
    448 elif using_copy_on_write():
    449     copy = False
--> 451 return self.apply(
    452     "astype",
    453     dtype=dtype,
    454     copy=copy,
    455     errors=errors,
    456     using_cow=using_copy_on_write(),
    457 )

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/m
anagers.py:352, in BaseBlockManager.apply(self, f, align_keys, **kwar
gs)
    350         applied = b.apply(f, **kwargs)
    351     else:
--> 352         applied = getattr(b, f)(**kwargs)
    353     result_blocks = extend_blocks(applied, result_blocks)
    355 out = type(self).from_blocks(result_blocks, self.axes)

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/internals/b
locks.py:511, in Block.astype(self, dtype, copy, errors, using_cow)
    491 """
    492 Coerce to the new dtype.
    493
   (...)
    507 Block
    508 """
    509 values = self.values
--> 511 new_values = astype_array_safe(values, dtype, copy=copy, erro
rs=errors)
    513 new_values = maybe_coerce_values(new_values)
    515 refs = None

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/asty
pe.py:242, in astype_array_safe(values, dtype, copy, errors)
    239     dtype = dtype.numpy_dtype
    241 try:
--> 242     new_values = astype_array(values, dtype, copy=copy)
    243 except (ValueError, TypeError):
    244     # e.g. _astype_nansafe can fail on object-dtype of string
s
    245     #  trying to convert to float
    246     if errors == "ignore":

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/asty
pe.py:187, in astype_array(values, dtype, copy)
    184     values = values.astype(dtype, copy=copy)
    186 else:
--> 187     values = _astype_nansafe(values, dtype, copy=copy)
    189 # in pandas we don't store numpy str dtypes, so convert to ob
ject
    190 if isinstance(dtype, np.dtype) and issubclass(values.dtype.ty
pe, str):

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/dtypes/asty
pe.py:138, in _astype_nansafe(arr, dtype, copy, skipna)
    134     raise ValueError(msg)
    136 if copy or is_object_dtype(arr.dtype) or
is_object_dtype(dtype):
    137     # Explicit copy, or required since NumPy can't view from
/ to object.
--> 138     return arr.astype(dtype, copy=True)
    140 return arr.astype(dtype, copy=copy)

ValueError: invalid literal for int() with base 10: '?'
```

And, argh, we get an error! If we look at the bottom of the error message, it seems that the error involves question marks ("?") in the data, which would also explain why this column imported as text rather than numbers in the first place.

Let's check.

In the cell below, use logical indexing to show the rows of `bcd` in which `bcd[bare_nuclei]` contains a question mark.

In [41]:
```
bcd['bare_nuclei']
```

Out[41]:
```
0       1
1      10
2       2
3       4
4       1
       ..
694     2
695     1
696     3
697     4
698     5
Name: bare_nuclei, Length: 699, dtype: object
```

Sure enough. Rather than leaving the cells of missing values empty, somebody has made the poor decision to enter question marks instead.

When you are dealing with other peoples' data, you'll find that this sort of the happens a LOT. It can be very aggravating, so we need to learn to treat these things as challenging puzzles instead of hassles!

Let's replace the question marks with nothing, so that this column becomes consistent with the rest. Fortunately, `DataFrame` (and `Series`) objects have a `replace()` function built in, so let's use that.

In [43]:
```
bcd['bare_nuclei'] = bcd['bare_nuclei'].replace('?', '')
```

In the cell below, confirm that we no longer have question marks in our "bare_nuclei" column.

In [45]:
```
bcd['bare_nuclei']
```

Out[45]:
```
0       1
1      10
2       2
3       4
4       1
       ..
694     2
695     1
696     3
697     4
698     5
Name: bare_nuclei, Length: 699, dtype: object
```

Note: As mentioned above, extracting columns or other subsets of data from a pandas `DataFrame` or `Series` does not create a new object but rather a new label to the existing object.

So, for example, `the_IDs = bcd['patient_id']` does not make a new object, but rather creates a second label referring to the original object (consistent with the behavior of base Python).

In general, however, pandas methods (functions) *do* create new objects. Thus, the step of assigning the output of `.replace()` back to the original data frame column is necessary.

In the cells below, confirm that the output of `.replace()` and `bcd['bare_nuclei']` have different IDs.

In [52]:
```
print("id of bcd['bare_nuclei']:", id(bcd['bare_nuclei']))
```

```
id of bcd['bare_nuclei']: 4963213456
```

In [55]:
```
print("ID of replaced_bare_nuclei:", id(replaced_bare_nuclei))
```

```
ID of replaced_bare_nuclei: 4963212560
```

And now we can convert the column to numeric values.

```
In [56]: bcd['bare_nuclei'] = pd.to_numeric(bcd['bare_nuclei'])
```

In the cell below, check the data types of columns in `bcd`.

```
In [57]: bcd.dtypes
```

```
Out[57]: patient_id                 int64
clump_thickness          float64
cell_size_uniformity     float64
cell_shape_uniformity      int64
marginal_adhesion          int64
single_ep_cell_size        int64
bare_nuclei              float64
bland_chromatin          float64
normal_nucleoli          float64
mitoses                    int64
class                     object
doctor_name               object
dtype: object
```

Okay! We have now have gotten our data somewhat into shape, meaning:

- missing data are actually missing
- columns of numeric data are numeric in type
- the column of doctor names contains only last names

So now we can explore some ways to deal with missing values.

Dealing with missing data

Finding missing values

Even though this dataset isn't all that large:

In [58]:
```
bcd.shape
```

Out[58]:
```
(699, 12)
```

699 rows is lot to look through "by hand" in order to find missing values.

We can test for missing values using the `DataFrame.isna()` method.

In [59]:
```
bcd.isna()
```

Out[59]:

	patient_id	clump_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adhesion
0	False	False	False	False	False
1	False	False	False	False	False
2	False	False	False	False	False
3	False	False	False	False	False
4	False	False	False	False	False
...	...	...	...	...	...
694	False	False	False	False	False
695	False	False	False	False	False
696	False	False	False	False	False
697	False	False	False	False	False
698	False	False	False	False	False

699 rows × 12 columns

By itself, that doesn't help us much. But if we combine it with summation (remember that `True` values count as 1 and `False` counts as zero):

In [60]:
```
bcd.isna().sum()
```

Out[60]:
```
patient_id              0
clump_thickness         1
cell_size_uniformity    1
cell_shape_uniformity   0
marginal_adhesion       0
single_ep_cell_size     0
bare_nuclei            18
bland_chromatin         4
normal_nucleoli         1
mitoses                 0
class                   0
doctor_name             0
dtype: int64
```

Now we have the counts by variable, and can easly see that there are missing values for a few of the variables.

Let's check some of the rows with missing values and make sure everything else looks normal in those rows. Notice above that the output of `.isna()` is Boolean, so we can use it to do logical indexing.

In [61]:
```
bcd[bcd['bland_chromatin'].isna()]
```

Out[61]:

	patient_id	clump_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adhesion
342	814265	2.0	1.0	1	1
343	814911	1.0	1.0	1	1
359	873549	10.0	3.0	5	4
365	897172	2.0	1.0	1	1

Okay, it looks like all of the other columns look fine.

In the cell below, check the rows that have missing values for either clump thickness or cell size uniformity. Do this in one go rather than separately (remember about the element-wise or operator, "|".

In [64]:
```
missing_values_rows = bcd[bcd['clump_thickness'].isnull() | bcd['cell_
print(missing_values_rows)
```

```
   clump_thickness  cell_size_uniformity
0              1.0                   NaN
2              NaN                   3.0
3              4.0                   NaN
```

So far so good. It looks like the rows that have missing values just have one missing value, and everything else seems fine. But let's do check that no rows have more than one missing value.

To do this, we can sum the number of missing values across the columns (i.e. within each row), and then see what the maximum number of missing values within a row is.

In [65]:
```
row_na_totals = bcd.isna().sum(axis = 1)
row_na_totals.max()
```

Out[65]:
```
1
```

So we see that no row has more than one missing value.

In the cell below, do the above calculation in one line.

In [67]:
```
row_na_max = bcd.isna().sum(axis=1).max()
print(row_na_max)
```

```
1
```

Dealing with missing values

Now that we have determined that there are missing values, we have to determine how to deal with them.

Ignoring missing values elementwise

One way to handle missing values is just to ignore them. Most of the standard math and statisitical functions will do that by default.

So this:

```
In [68]: bcd['clump_thickness'].mean()
Out[68]: 3.0
```

Computes the mean clump thickness ignoring the one missing value.

We can compute the mean (again ignoring missing values) for all the numeric columns like this:

```
In [69]: bcd.mean(numeric_only = True)  # the numeric_only refers to columns, r
Out[69]: clump_thickness         3.000000
         cell_size_uniformity    3.333333
         dtype: float64
```

That worked, but the output is a little awkward because the patient ID is being treated as a numeric variable. We can fix that by converting the patient ID variable to a string variable.

```
In [70]: bcd['patient_id'] = bcd['patient_id'].astype('string')

---------------------------------------------------------------------------
KeyError                                  Traceback (most recent call
last)
File ~/anaconda3/lib/python3.11/site-packages/pandas/core/indexes/bas
e.py:3653, in Index.get_loc(self, key)
   3652 try:
-> 3653     return self._engine.get_loc(casted_key)
   3654 except KeyError as err:

File ~/anaconda3/lib/python3.11/site-packages/pandas/_libs/index.pyx:
147, in pandas._libs.index.IndexEngine.get_loc()
```

```
File ~/anaconda3/lib/python3.11/site-packages/pandas/_libs/index.pyx:
176, in pandas._libs.index.IndexEngine.get_loc()

File pandas/_libs/hashtable_class_helper.pxi:7080, in pandas._libs.ha
shtable.PyObjectHashTable.get_item()

File pandas/_libs/hashtable_class_helper.pxi:7088, in pandas._libs.ha
shtable.PyObjectHashTable.get_item()

KeyError: 'patient_id'

The above exception was the direct cause of the following exception:

KeyError                                  Traceback (most recent call
last)
Cell In[70], line 1
----> 1 bcd['patient_id'] = bcd['patient_id'].astype('string')

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/frame.py:37
61, in DataFrame.__getitem__(self, key)
   3759 if self.columns.nlevels > 1:
   3760     return self._getitem_multilevel(key)
-> 3761 indexer = self.columns.get_loc(key)
   3762 if is_integer(indexer):
   3763     indexer = [indexer]

File ~/anaconda3/lib/python3.11/site-packages/pandas/core/indexes/bas
e.py:3655, in Index.get_loc(self, key)
   3653     return self._engine.get_loc(casted_key)
   3654 except KeyError as err:
-> 3655     raise KeyError(key) from err
   3656 except TypeError:
   3657     # If we have a listlike key, _check_indexing_error will r
aise
   3658     #  InvalidIndexError. Otherwise we fall through and re-ra
ise
   3659     #  the TypeError.
   3660     self._check_indexing_error(key)

KeyError: 'patient_id'
```

And now the means should look a little better because we won't have the mean for the ID column in the millions>

Recompute the mean for the numeric columns in the cell below.

```
In [ ]:
```

Removing missing values

We are about to start learning how to remove missing values from our data frame, *however...*

Before we start messing around too much with the values in our data frame, let's make sure we can easily "hit the reset button" and get back to a nice starting point. To do this, we'll want to

- reload the data
- modify the column of Dr. names
- set the patient ID to type str
- remove the question marks from the bare nuclei column
- set the bare nuclei column to numeric

This is a perfect job for a function!

In the cell below, finish writing the function to reset our data frame to the desired starting point.

```
In [71]: def hit_reset():
             bcd = pd.read_csv('./data/breast_cancer_data.csv')
             bcd['patient_id'] = ...
             ...

             return bcd
```

Removing rows with missing values

Obviously, rows in which all values are missing won't do us any good, so we can drop them with:

```
In [72]: bcd = bcd.dropna(how = 'all')
```

This drops rows in which *all* of the values are missing. This code ran without error, but we know it also didn't do anything in this case because we don't have any rows in which all the values are missing!

Sometimes a case can be made for throwing out all observations (rows) that are incomplete, that is, if they contain *any* missing values.

```
In [73]: bcd = bcd.dropna(how = 'any')
```

In the cell below, check the (new) shape of `bcd` .

```
In [74]: bcd.shape
```

```
Out[74]: (2, 2)
```

It should have fewer rows now.

And now is a perfect time to test our function! In the cell below, hit the reset button on bcd.

In [79]:
```
bcd = pd.read_csv('./breast_cancer_data.csv')
bcd
```

Out[79]:

	patient_id	clump_thickness	cell_size_uniformity	cell_shape_uniformity	marginal_adhesion
0	1000025	5.0	1.0	1	1
1	1002945	5.0	4.0	4	5
2	1015425	3.0	1.0	1	1
3	1016277	6.0	8.0	8	1
4	1017023	4.0	1.0	1	3
...	...	...	...	...	...
694	776715	3.0	1.0	1	1
695	841769	2.0	1.0	1	1
696	888820	5.0	10.0	10	3
697	897471	4.0	8.0	6	4
698	897471	4.0	8.0	8	5

699 rows × 12 columns

Check the shape.

In [81]:
```
bcd.shape
```

Out[81]:
```
(699, 12)
```

Check the data types of the columns.

In [82]:
```
bcd.dtypes
```

Out[82]:
```
patient_id                  int64
clump_thickness           float64
cell_size_uniformity      float64
cell_shape_uniformity       int64
marginal_adhesion           int64
single_ep_cell_size         int64
bare_nuclei                object
bland_chromatin           float64
normal_nucleoli           float64
mitoses                     int64
class                      object
doctor_name                object
dtype: object
```

Check the doctor name column.

In [85]:
```
doc= bcd['doctor_name']
doc
```

Out[85]:
```
0        Dr. Doe
1      Dr. Smith
2        Dr. Lee
3      Dr. Smith
4       Dr. Wong
         ...
694      Dr. Lee
695    Dr. Smith
696      Dr. Lee
697      Dr. Lee
698     Dr. Wong
Name: doctor_name, Length: 699, dtype: object
```

Removing columns with missing values

And we could do the same for columns if we wished, though this is less frequently done. We just need to change the axis (direction) over which `DataFrame.dropna()` works.

In [86]:
```
bcd = bcd.dropna(axis = 1, how = 'any') # drop columns rather than row
```

This leaves us with only the complete columns.

In [87]:
```
bcd.shape
```

Out[87]:
```
(699, 7)
```

Let's see which they are.

In [88]:
```
bcd.columns
```

Out[88]:
```
Index(['patient_id', 'cell_shape_uniformity', 'marginal_adhesion',
       'single_ep_cell_size', 'mitoses', 'class', 'doctor_name'],
      dtype='object')
```

Filling in missing values

Occasionally, we may want to fill in missing values. This isn't very common, but might be useful if some other function you are using doesn't handle missing values gracefully.

Before filling in missing values, we need to restore our data frame so it actually has missing values. Good thing we wrote that function!

In [108]:
```
bcd = hit_reset()
```

```
   1668     errors=self.options.get("encoding_errors", "strict"),
   1669     storage_options=self.options.get("storage_options", None)
,
   1670 )
   1671 assert self.handles is not None
   1672 f = self.handles.handle

File ~/anaconda3/lib/python3.11/site-packages/pandas/io/common.py:859
, in get_handle(path_or_buf, mode, encoding, compression, memory_map,
is_text, errors, storage_options)
    854 elif isinstance(handle, str):
    855     # Check whether the filename is to be opened in binary mo
de.
    856     # Binary mode does not support 'encoding' and 'newline'.
    857     if ioargs.encoding and "b" not in ioargs.mode:
    858         # Encoding
--> 859         handle = open(
    860             handle,
    861             ioargs.mode,
```

In [109]:
```
bcd
```

Out[109]:

	patient_id	cell_shape_uniformity	marginal_adhesion	single_ep_cell_size	mitoses	clas
0	1000025	1	1	2	1	benig
1	1002945	4	5	7	1	benig
2	1015425	1	1	2	1	benig
3	1016277	8	1	3	1	benig
4	1017023	1	3	2	1	benig
...	...	...	...	...	...	..
694	776715	1	1	3	1	benig
695	841769	1	1	2	1	benig
696	888820	10	3	7	2	malignan
697	897471	6	4	3	1	malignan
698	897471	8	5	4	1	malignan

699 rows × 7 columns

We can fill in missing values with any single value we want, such as a zero.

```
In [90]: bcd = bcd.fillna(0)
```

In the cell below, check to see that we no longer have missing values.

```
In [94]: missing = bcd.isnull().sum()
         print(missing)
```

```
patient_id              0
cell_shape_uniformity   0
marginal_adhesion       0
single_ep_cell_size     0
mitoses                 0
class                   0
doctor_name             0
dtype: int64
```

In the cell below, reset the data and verify that the missing data are back.

```
In [96]: print(bcd.head())
```

```
   patient_id  cell_shape_uniformity  marginal_adhesion  single_ep_ce
ll_size  \
0     1000025                      1                  1
2
1     1002945                      4                  5
7
2     1015425                      1                  1
2
3     1016277                      8                  1
3
4     1017023                      1                  3
2

   mitoses   class doctor_name
0        1  benign     Dr. Doe
1        1  benign   Dr. Smith
2        1  benign     Dr. Lee
3        1  benign   Dr. Smith
4        1  benign    Dr. Wong
```

In the cell below, fill the missing values in each column with the column mean. (Hint: this is pandas, so this is actually easy!)

In [107]:
```
print(bcd_filled.head())
```

```
   patient_id  cell_shape_uniformity  marginal_adhesion  single_ep_ce
ll_size  \
0     1000025                      1                  1              
2   
1     1002945                      4                  5              
7   
2     1015425                      1                  1              
2   
3     1016277                      8                  1              
3   
4     1017023                      1                  3              
2   

   mitoses   class doctor_name  
0        1  benign     Dr. Doe  
1        1  benign   Dr. Smith  
2        1  benign     Dr. Lee  
3        1  benign   Dr. Smith  
4        1  benign    Dr. Wong  
```

And now verify that there are no more missing values.

In [100]:
```
missing_values_count = bcd.isnull().sum()
print(missing_values_count)
```

```
patient_id               0
cell_shape_uniformity    0
marginal_adhesion        0
single_ep_cell_size      0
mitoses                  0
class                    0
doctor_name              0
dtype: int64
```

Summary

In this tutorial, we learned or remembered how to do some of the foundational data wrangling tasks. These are:

- importing data into pandas from a data file
- cleaning up the data in the columns
- converting columns to the appropriate type
- removing or filling in missing values